DISS. ETH NO. 25751

Identification of causal mechanisms of cell variability in signaling networks from single-cell time series snapshots

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

ANNA KLIMOVSKAIA

Specialist in Mathematics and System Programming
Lomonosov Moscow State University
born on August 2, 1990
citizen of Smolensk, Russia

accepted on the recommendation of

Prof. Dr. Manfred Claassen
Prof. Dr. Uwe Sauer
Prof. Dr. Nicolai Felix Meinshausen

2019

Contents

Abstract vi

Zusammenfassung x

Part I Introduction and Background

1. Introduction 1
    - 1.0.1. Mechanistic modeling 3
    - 1.0.2. Statistical approaches for hypothesis generation 4
    - 1.0.3. Thesis contributions 6

Part II Scientific Contributions

2. Sparse regression based structure learning of stochastic reaction networks from single cell snapshot time series 13
    - 2.1. Introduction 14
    - 2.2. Results 15
        - 2.2.1. Sparse regression for structure learning of stochastic reaction networks 15
        - 2.2.2. Reactionet lasso 17
        - 2.2.3. Ab initio structure learning of chemical reaction networks 20
        - 2.2.4. Structure learning of large chemical reaction networks with prior knowledge 26
    - 2.3. Discussion 28

2.4. Methods . . . 31
  2.4.1. Experimental setting: single cell time series snapshot data . . . 31
  2.4.2. Moment equations for the Chemical Master Equation . . . 31
  2.4.3. Gradient matching for parameter estimation of ordinary differential equation systems . . . 32
  2.4.4. Moment estimation for noisy single cell data . . . 33
2.5. Acknowledgments . . . 34

3. Identification of mechanisms of fractional killing in TRAIL-induced apoptosis from mass cytometry time series snapshots 37
3.1. Introduction . . . 38
3.2. Results . . . 40
  3.2.1. Analysis of TRAIL-induced Apoptosis by Mass Cytometry . . . 40
  3.2.2. Inference of asynchrony inducing mechanisms from single-cell snapshot time series with RealMatch . . . 42
  3.2.3. Identifying signaling nodes of survivor subpopulation after TRAIL Treatment . . . 45
  3.2.4. Identification of potential mechanisms of TRAIL-induced apoptosis regulated by the identified key nodes . . . 47
3.3. Conclusion . . . 49
3.4. Methods . . . 55
  3.4.1. Reagents . . . 55
  3.4.2. Clonogenic Survival Assays . . . 57
  3.4.3. Mass Cytometry . . . 57
  3.4.4. Computational Approach . . . 58

4. Causal learning of signaling pathways from single-cell time series snapshots 61
4.1. Introduction . . . 62
4.2. Results . . . 66
  4.2.1. Protein signaling as Causal Time Series Snapshots Model (CTSSM) 66
  4.2.2. Mass Cytometry Causal Reactions Annotation (MassCaRA) . . 68
  4.2.3. Experimental results and benchmarks . . . 70
4.3. Methods . . . 74
  4.3.1. Learning sparse skeleton . . . 74
  4.3.2. Orient edges: "soft" interventions with transient transfection . . 75

4.3.3. Simulating synthetic experiments 75
4.4. Discussion 77
4.5. Acknowledgements 78

Part III Concluding Remarks

5. Concluding Remarks **83**

6. Acknowledgements **89**

Appendices **91**

A. Supplementary material Chapter 2 **93**
A.1. S1 Text 93
A.2. S2 Text 95
A.3. S3 Text 96
A.4. S4 Text 97
A.5. S1 Dataset 98
A.6. Supplementary Figures 99

B. Supplementary material Chapter 3 **111**

C. Supplementary material Chapter 4 **113**
List of Figures 116

7. Curriculum vitae **123**

Abstract

Molecular systems biology constitutes a holistic approach towards understanding mechanisms of the cell functioning. Cell signaling is one of the examples of self regulation mechanisms to perform functional roles such as proliferation, stress or apoptosis. Aberrant signaling processes can lead to diseases such as cancer, autoimmunity, and diabetes. Therefore, understanding the mechanisms of signaling in general as well as specific changes in the context of the disease, is vital for the design of novel treatment strategies.

Heterogeneity of cancer in terms of genetic mutations is a well known cause of treatment resistance. Recent studies in cancer cell lines have demonstrated a crucial role of non-genetic origins of cancer heterogeneity, for example fractional killing of sister cells in TRAIL-induced apoptosis. Various modeling approaches have been proposed to investigate this phenomenon. However, apoptosis is a process with many components and non-trivial dynamics, which makes its modeling a particularly challenging task; in particular, to date no model has provided a satisfactory explanation of fractional killing.

Studying responses that affect only a part of an apparently homogeneous population - as in the case of fractional killing - requires monitoring cellular properties and characteristics at the single cell level. Recently developed multi-parametric high throughput technologies for single cell measurements, such as mass cytometry, allow the monitoring of 30+ proteins at the single cell level for up to millions of individual cells in a single experiment, and thereby enables us to learn detailed descriptions, i.e. models of this process. However, mass cytometry is a destructive technique, which therefore doesn't provide time series information, but rather time dynamics acquired by disjoint snapshots.

In this thesis we focus on computational approaches towards discovering causal non-genetic mechanisms of cell-to-cell variability in protein signaling from mass cytometry snapshots. In particular, we focus on mathematical models of cell variability and computational methods aimed to overcome two major limitations of mass cytometry to reverse engineer these models from data: (a) lack of temporal connection between the cells in different snapshots; (b) partially observed system due to a limited panel of measured markers.

In **Chapter 2** we present *Reactionet lasso*, a regression-based gradient matching approach for stochastic reaction networks, that is capable of partial structure learning for systems of arbitrary size. We assessed the structure learning capabilities of the *Reactionet lasso* on synthetic data for systems of different size and complexity. *Reactionet lasso* achieves structure learning for problem instances hundreds of orders of magnitude larger than previously reported. This approach opens the prospect of obtaining quantitative and predictive reaction models in many areas of biology and medicine, and in particular areas such as cancer biology, which are characterized by significant system alterations and many unknown reactions.

Chapter 3 demonstrates application of *Reactionet lasso* towards understanding molecular mechanisms of fractional killing in TRAIL-induced apoptosis. In this chapter we address both problems of destructive measurements and partially observed system by a combination of optimal transport and *Reactionet lasso* (*REALmatch*). *REALmatch* was applied to a mass cytometry time course study of TRAIL-induced apoptosis and derived mechanistic insights into fractional killing. The proposed mechanisms were evaluated through functional inhibitor experiments that perturb key pathways in identified network.

While the focus of chapters 2 and 3 is on the discovery of *specific* mechanisms of protein signaling, **Chapter 4** attempts to address a more general question of modeling of a signaling system as a whole. In particular, we raise a question of a necessity of acquiring a precise mechanistic model to answer a question of "why?" and "what if?" in the context of protein signaling. We propose to model signaling mechanisms with dynamic Structural Equations Models, and introduce *MassCaRA*, a simple structure learning approach to learn the model from single-cell mass cytometry snapshots. We demonstrate the performance *MassCaRA* for *ab initio* reconstruction of signaling pathway in the presence of latent variables.

To summarize, in this thesis we propose several different computational approaches towards understanding causal mechanisms of cell variability in protein signaling from mass cytometry snapshots. The proposed methods are generic and could be in principle applied to any signaling system. However, we suggest that *REALmatch* is better suitable for systems with a bi-modal outcome like TRAIL-induced apoptosis, while *MassCaRA* should give a better performance on the systems with a rather homogeneous behavior.

Zusammenfassung

Die Molekularsystembiologie stellt einen ganzheitlichen Ansatz zum Verständnis der Mechanismen der Zellfunktion dar. Die Zellsignalisierung ist eines der Beispiele für Selbstregulierungsmechanismen zur Erfüllung funktionaler Funktionen wie Proliferation, Stress oder Apoptose. Unregelmäßige Signalprozesse können zu Krankheiten wie Krebs, Autoimmunität und Diabetes führen. Daher ist das Verständnis der Mechanismen der Signalübertragung im Allgemeinen sowie spezifischer Veränderungen im Zusammenhang mit der Erkrankung von entscheidender Bedeutung für die Entwicklung neuer Behandlungsstrategien.

Die Heterogenität von Krebs im Hinblick auf genetische Mutationen ist eine bekannte Ursache für die Resistenz gegen die Behandlung von Krebs. Jüngste Studien an Krebszelllinien haben gezeigt, dass die nicht-genetische Herkunft der Krebsheterogenität eine entscheidende Rolle spielt, z.B. der Log-cell-kill von Schwesterzellen bei der TRAIL-induzierten Apoptose. Verschiedene Modellierungsansätze wurden vorgeschlagen, um dieses Phänomen zu untersuchen. Allerdings ist die Apoptose ein Prozess mit vielen Komponenten und nicht-trivialer Dynamik, was ihre Modellierung zu einer besonders anspruchsvollen Aufgabe macht; insbesondere hat bisher noch kein Modell eine zufriedenstellende Erklärung für den Log-cell-kill geliefert.

Die Untersuchung von Reaktionen, die nur einen Teil einer scheinbar homogenen Population betreffen - wie im Falle des Log-cell-kill - erfordert die Überwachung der zellulären Eigenschaften und Merkmale auf Einzelzellenebene. Neu entwickelte multiparametrische Hochdurchsatztechnologien für Einzelzellmessungen, wie z.B. die Massenzytometrie, ermöglichen die Überwachung von 30+ Proteinen auf Einzelzellenebene für bis zu Millionen von Einzelzellen in einem einzigen Experiment und ermöglichen es uns so, detaillierte Beschreibungen, d.h. Modelle dieses Prozesses zu erlernen. Die Massenzytometrie ist jedoch eine zerstörerische Technik, die daher keine Zeitreiheninformationen liefert, sondern Zeitdynamiken, die durch disjunkte Schnappschüsse gewonnen werden.

In dieser Arbeit konzentrieren wir uns auf computergestützte Ansätze zur Entdeckung

kausaler nicht-genetischer Mechanismen der Zell-zu-Zell-Variabilität in der Proteinsignalisierung aus Schnappschüssen der Massenzytometrie. Insbesondere konzentrieren wir uns auf mathematische Modelle der Zellvariabilität und Berechnungsmethoden, die darauf abzielen, zwei große Einschränkungen der Massenzytometrie zu überwinden, um diese Modelle aus Daten zu rekonstruieren: (a) fehlende zeitliche Verbindung zwischen den Zellen in verschiedenen Momentaufnahmen; (b) teilweise beobachtetes System aufgrund einer begrenzten Anzahl von gemessenen Markern.

In **Kapitel 2** stellen wir *Reactionet lasso* vor, einen regressionsbasierten Gradientenabgleichsansatz für stochastische Reaktionsnetzwerke, der in der Lage ist, strukturiertes Lernen für Systeme beliebiger Größe durchzuführen. Wir bewerteten die strukturierten Lernfähigkeiten des *Reactionet Lasso* an synthetischen Daten für Systeme unterschiedlicher Größe und Komplexität. *Reactionet lasso* erreicht strukturiertes Lernen für Probleme, die Hunderte von Größenordnungen größer sind als bisher berichtet. Dieser Ansatz eröffnet die Möglichkeit, quantitative und prädiktive Reaktionsmodelle in vielen Bereichen der Biologie und Medizin zu erhalten, insbesondere in der Krebsbiologie, die sich durch signifikante Systemänderungen und viele unbekannte Reaktionen auszeichnen.

Kapitel 3 demonstriert die Anwendung von *Reactionet lasso* zum Verständnis der molekularen Mechanismen des des Log-cell-kill bei der TRAIL-induzierten Apoptose. In diesem Kapitel behandeln wir sowohl Probleme zerstörerischer Messungen als auch teilweise beobachtete Systeme durch eine Kombination aus optimalem Transport und *Reactionet lasso* (*REALmatch*). *REALmatch* wurde auf eine massenzytometrische Zeitverlaufsstudie zur TRAIL-induzierten Apoptose angewendet und daraus mechanistische Erkenntnisse über den Log-cell-kill abgeleitet. Die vorgeschlagenen Mechanismen wurden durch funktionelle Inhibitorexperimente evaluiert, die Schlüsselwege im identifizierten Netzwerk stören.

Während der Schwerpunkt der Kapitel 2 und 3 auf der Entdeckung von *spezifischen* Mechanismen der Proteinsignalisierung liegt, versucht **Kapitel 4**, eine allgemeinere Frage der Modellierung eines Signalsystems als Ganzes zu behandeln. Insbesondere stellen wir die Frage nach der Notwendigkeit, ein präzises mechanistisches Modell zu erwerben, um die Frage nach “Warum?” und “Was, Wäre, Wenn?” im Zusammenhang mit der Proteinsignalisierung zu beantworten. Wir schlagen vor, Signalmechanismen mit dynamischen Strukturgleichungsmodellen zu modellieren und stellen *MassCaRA* vor, einen einfachen strukturierten Lernansatz, um das Modell aus Einzelzell-Massenzytometrie-Schnappschüssen zu lernen. Wir demonstrieren die Ergebnisse von *MassCaRA* für die *ab initio* Rekonstruktion des Signalwegs in Gegenwart eines Konstrukts.

Zusammenfassend lässt sich sagen, dass wir in dieser Arbeit mehrere verschiedene Berechnungsansätze zum Verständnis der kausalen Mechanismen der Zellvariabilität in der Proteinsignalisierung aus Schnappschüssen der Massenzytometrie vorschlagen. Die vorgeschlagenen Methoden sind generisch und könnten prinzipiell auf jedes Signalisierungssystem angewendet werden. Wir schlagen jedoch vor, dass *REALmatch* besser für Systeme mit einem bimodalen Ergebnis wie TRAIL-induzierte Apoptose geeignet ist, während *MassCaRA* eine bessere Leistung auf den Systemen mit einem eher homogenen Verhalten bieten sollte.

Part I.

Introduction and Background

1 Introduction

> *We have a large reservoir of engineers (and scientists) with a vast background of engineering know how. They need to learn statistical methods that can tap into the knowledge. Statistics used as a catalyst to engineering creation will, I believe, always result in the fastest and most economical progress...*
>
> – George Box, 1992

The relatively young field of Systems Biology emerged from the concept of interpreting a cell as a "complex system". Essentially Systems Biology aims at studying how relationships between a system's parts give rise to the collective behaviour of a cell (e.g. proliferation, apoptosis, etc.). Another important question addressed by Systems Biology is how changes in the systems components lead to a aberrant behavior such as disease. In contrast, classical molecular biology is centered around gathering "facts" about the cell such as genome sequences, protein properties, etc. However, these facts alone are not enough to understand the complex behavior. The common analogy in systems biology is with a car: knowing every detail about of the component parts of the car will not help you to construct a working vehicle, unless you know how they should be assembled together. Even then, in order to make it drive, you also need to know which "button to turn on". One of the most important applications of understanding molecular biology at a systems level is drug discovery and treatment optimization [1, 2], which in the car analogy is this "button".

State of the art research in systems biology clearly distinguishes two modeling approaches: (1) knowledge discovery (data-mining) and (2) simulation of the system from a mechanistic model to predict its behavior and potentially validate the prediction with experiments [2]. Mechanistic approaches (Section 1.0.1) constitute the most comprehensive modeling approach to understand systems behavior. However, typically the structure of mechanistic models is assumed to be known. When this is not the case, data-mining approaches can be helpful. Data-mining approaches are typically centered around statistical methods (Section 1.0.2) and attempt to discover new potential causal interactions not necessarily connecting them into a unified network. We discuss advantages and challenges of both approaches, and their role in systems modeling, in the sections below.

Systems biology research and in particular its computational approaches are largely driven by data. The majority of methods to quantify components of the cell (e.g. mRNA, proteins) rely on bulk measurements. Therefore, historically, since the measured data was dictating the modeling approach, the majority of research efforts were invested into modeling deterministic dynamics systems, by means of Ordinary Differential Equations [3]. However, these models could only capture the average behavior of cells in a heterogeneous population. Development of single-cell technologies such as flow cytometry allowed to unlock a new layer of modeling: quantitatively assessing heterogeneity. Stochasticity was shown to play a crucial role in all types of molecular systems from stochastic gene expression [4–6] to signaling networks [7].

The development of mass cytometry was one of the significant "break-throughs" in addressing the problem of modeling cell-to-cell variability [8, 9]. Mass cytometry is a single cell proteomic technique that expands conventional flow cytometry approaches, permitting of multi-parametric monitoring of 30+ proteins at the single cell level for up to millions of individual cells in a single experiment. One drawback of this technology is that like any cytometry measurement technique it is destructive. Temporal information about the processes is therefore only partially preserved by measuring snapshots at different times after stimulation.

From the modeling point of view there are two major sources of cell-to-cell variability: *intrinsic* and *extrinsic* [10]. Intrinsic variability is the noise due to the stochastic nature of chemical kinetics. In mechanistic modeling this source of noise is modeled with Chemical Master Equation by means of Stochastic Differential Equations [11]. Extrinsic noise it typically referred to as heterogeneity in the cell population, such as variations in total protein concentration, in cell volume and cell cycle stage [12]. In terms of mechanistic modeling this source of variability could be modeled in two ways: random initial conditions and variability reaction rates.

1.0.1. Mechanistic modeling

Mechanistic modeling was the state of the art long before the development of single-cell technologies. It has its roots in Control Systems theory, and provides a comprehensive toolkit to study systems [13, 14]: their structure, dynamics, control and design.

Systems structure. The first step in mechanistic modeling is defining network structure, such as gene regulatory networks or signaling pathway diagrams. The development of a unified language for representing a structure of a biological network (Systems Biology Markup Language) facilitated a rich collection of gene interaction networks and biochemical pathways [15], making them easily accessible. Many software were developed to read out the models into graphical format (diagram), or to simulate the corresponding dynamic system.

Given a known model one of the main problems in mechanistic modeling is parameter estimation. While parameter estimation from time-series experiments of deterministic systems is already a standard procedure [16], efficient modeling of biological stochastic systems is still a challenge, due to the destructive nature of cytometry experiments [17, 18]. However, an efficient approach was proposed for single-cell time lapse data [5]. Another challenge in parameter inference in general is practical identifiability, sensitivity of the parameter with respect to the observed data, which may limit identifiability of the model from data [19].

In the case when the model structure is not known, one might want to reverse engineer it from data. Here the state of the art approach is model selection: several plausible structures are enumerated, fit to data, and the best fit is chosen [20, 21]. Though this approach is the most widely used among biologists, it is not scalable to large systems (>20 components) and many unknown interactions.

Systems design. Optimal experiment design has been demonstrated to successfully address the problem of practical identifiability for deterministic [22, 23] and stochastic systems [24, 25]. However, topological identifiability of an arbitrary large system still remains an open question [26].

Systems dynamics. Knowing the structure of the model the next step in the systems analysis typically includes the study of dynamics. One way to understand the behavior of the system is to study its sensitivity under changes of parameters [27, 28]. Typically this problem is addressed by simulations. Simulation of a deterministic biological system directly from an SBML file is now a part of routine pipeline and incorporated in the majority of standard software. Simulations of stochastic systems are more complicated in their nature. The Stochastic Simulations Algorithm [29] is the state of the art algorithm for simulation of a stochastic reaction network. However, its performance might be too slow for large scale dynamic networks. A number of fast approximations were recently proposed: tau-leaping [30], finite state projection algorithm [31], linear noise approximation [32].

Systems control. Modeling drug discovery is largely based on principles from control theory for deterministic systems. Advances were also made in the control of stochastic systems: "in silico" feedback control was configured to achieve precise and robust set point regulation of gene expression [33].

The described properties are highly desirable for systems biology modeling, but are largely precluded by the complexity of structure learning.

1.0.2. Statistical approaches for hypothesis generation

The second pillar of computational systems biology is data mining: the application of statistical and machine learning methods to gather facts from data for complex systems. Development of the aforementioned high-throughput single-cell technologies such as flow and mass cytometry, opened many opportunities for transferring statistical and machine learning algorithms to biological applications. In particular, statistical approaches come in hand for a difficult ill-posed problem of learning topology of a dynamic system: there were numerous attempts to exploit statistical concepts to generate hypothesis about chemical reactions. Below we give an overview on causal structure learning from snapshot data. As we dismissed in Section 1.0.1, one of the caveats of cytometry techniques is the lack of temporal connection between the snapshots; detailed discussion of causal discovery from time series [34, 35] is therefore beyond the scope of this thesis.

Probabilistic graphical models [36] constitute a popular choice when attempting to generate hypothesis about biochemical reactions. Friedman first proposed to analyze gene expression data from DNA microarrays using Bayesian network[37]. Later Sachs et al. [38] showed promising results on the application of Bayesian Networks towards signaling. Bayesian Networks represent conditional dependencies and independencies in the data. In some cases directions of edges in Bayesian Networks coincide with true causal directions, however this is not true in a general case [39].

Extension of Bayesian Networks to tackle causality problem was proposed by means of

Causal Bayesian Networks [39]. The notion of *causation* is defined in terms of interventions: if A is causal to B, intervening on A would lead to an effect on B. Therefore, the theory of causal networks relies on learning from both observational and interventional data. In case when only observational data is available, conclusions about directions are limited to the Markov equivalence class of models: only some edges can be directed while orientation of the rest remains ambiguous [40]. To resolve the remaining edges, interventional experiments are needed. In biological settings, such interventions are gene knockouts for gene regulatory networks [41] inference or molecular inhibitors for signaling systems[42]. Recently, a new type of interventions became available for signaling systems: interventions by transient transfection [43].

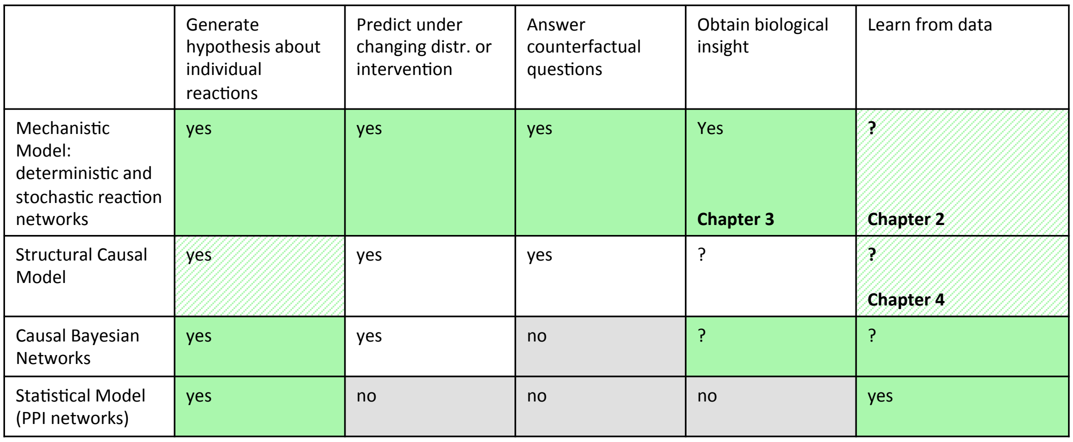
While causal inference in the presence of interventional data demonstrated significant improvement, it is still limited by the assumption of a network to be a directed acyclic graph. The Sparse Bayesian algorithm, a Byaesian modeling approach of a steady-state distribution of a dynamic system, was proposed to overcome this limitation to infer gene regulatory interactions from gene expression microarrays and gene-knockout interventions[44].

Sachs et al. [45] first raised the question of the importance of dynamics in statistical modeling from single-cell cytometry snapshots data. All the models described above assume that the system is at a steady-state. But as Sachs pointed out, a signaling system is never truly in a steady state. The authors demonstrate how to use Generalized Bayesian networks (GBNs)[46, 47], a generalization of Bayesian Networks to accommodate cycles, coupled with small molecular inhibitors.

Recently Triantafillou et al. demonstrated how advances in causal structure learning methods from interventional data, under the assumption of causal Bayesian Networks, could be applied to mass cytometry data [48]. Based on the results of their study, the authors once again stress that noise, confounders and feedback cycles are still open questions for the successful application of automatic causal discovery towards biology.

Another approach for causal modeling was recently proposed by using Structural Equation Models (SEMs) [49]. Using large-scale gene deletion experiments in yeast, the authors aim to predict the effect of a new gene knockout or knockdown on a phenotype of interest. Compared to the previously described models, this is already a first step towards bridging statistical models with comprehensive understanding of system's behaviour. Bongers et al. demonstrated relation of steady-state of a dynamic system and SEMs [50]. Although this theoretical finding did not show practical applications, yet we consider it as an important step towards bridging statistical models and mechanistic ones.

1.0.3. Thesis contributions



	Generate hypothesis about individual reactions	Predict under changing distr. or intervention	Answer counterfactual questions	Obtain biological insight	Learn from data
Mechanistic Model: deterministic and stochastic reaction networks	yes	yes	yes	Yes **Chapter 3**	**?** **Chapter 2**
Structural Causal Model	yes	yes	yes	?	**?** **Chapter 4**
Causal Bayesian Networks	yes	yes	no	?	?
Statistical Model (PPI networks)	yes	no	no	no	yes

Figure 1.1.: A simple taxonomy of models from Peters et al. in the light of modern Systems Biology (proteomics) research and contribution of this thesis.

Green boxes represent modeling approaches covered by systems biology applications, "yes/no/?" correspond to the overall state of the art in modeling complex systems beyond biology. Light green boxes represent recent progress of these approaches towards systems biology applications, but without consistently demonstrated impact.

Central questions in systems biology are "Why?", "How?" and "What if?". The main focus of this thesis is the investigation of mathematical modeling approaches, for answering these questions for the scenario where the network structure of the systems is unknown. In particular, we are interested in recovering non-genetic mechanisms of cell-to-cell variability. Figure 1 depicts a summary table of modern approaches for modeling complex systems [51] in the light of modern Systems Biology research (in the context of protein signaling) and the contribution of this thesis. Starting from classical dynamic systems formulations, the work dives deeper into statistical modeling. Like George Box we believe that statistical modeling could be invaluable towards understanding of large-scale complex systems.

Chapter 2 introduces *Reactionet lasso*, a computational procedure that derives a step-wise sparse regression approach on the basis of the Chemical Master Equation, enabling large-scale structure learning for reaction networks by implicitly accounting for billions of topology variants. We have assessed the structure learning capabilities of the Reactionet lasso on synthetic data for the complete TRAIL induced apoptosis signaling cascade comprising 70 reactions. We find that the Reactionet lasso is able to efficiently recover the structure of these reaction systems, ab initio, with high sensitivity and specificity. With only $< 1\%$ false discoveries, the Reactionet lasso is able to recover 45% of all true reactions ab initio among > 6000 possible reactions and over 10^{2000} network topologies. We provide software to allow for wide applicability of the Reactionet lasso.

Chapter 3 addresses the problem of studying asynchronous processes from mass cytometry snapshots. One of the limitations of flow and mass cytometry is its destructive nature, making it difficult to apply to problems when the changes in the systems behavior appear over time and could happen at different speed for individual cells. Non-genetic origins of fractional killing during TRAIL-induced apoptosis is a perfect example of a problem where knowledge of temporal connections between the snapshots is vital: starting from a seemingly homogeneous population cells show differences in time of committing death as well as appearance of a resistant population by the end of the process. We introduce *RealMatch*, a computational approach based on *Reactionet lasso* coupled with optimal transport to generate hypothesis about molecular mechanisms of fractional killing in TRAIL-induced apoptosis.

Chapter 4 aims at filling in the gap between mechanistic and statistical models. Structure and parameter learning for dynamic systems still remains a challenging task. Recently progress has been made in linking statistical concept of distribution and correlation, to a more philosophical notion of "causation". Although everyone knows the mantra "correlation doesn't apply causation" a lot of effort in the last decade was introduced towards mathematical formalism of extracting "causal" direction and exploiting this knowledge [52][51]. We propose Causal Time Series Snapshots model (CTSSM), a simple statistical model based on causality principles, which extend classical statistical models towards modeling essential biological concepts as feedback loops. This model belongs to a Structural Equations Models class and allows to apply all the respective formalism developed in causal inference and counterfactual analysis. In the context of biology it means that it could be used to quantitatively answer questions such as "given a certain intervention on the system (e.g. with a drug), how will it affect the phenotype?". Compared to mechanistic models, CTSSM avoids tedious computations of differential equations by approximating them with functional dependencies. We introduce *MassCaRA*, a simple heuristic based approach to learn the structure of CTTSM and demonstrate its performance on several synthetic datasets including one generated from a mechanistic model. The ability of CTSSM to model feedback loops and the structure learning capabilities of *MassCaRA* make CTSSM an attractive modeling tool for systems biology applications.

References

[1] L. Hood, J. R. Heath, M. E. Phelps, and B. Lin. "Systems biology and new technologies enable predictive and preventative medicine". *Science* **306**:5696 (2004), pp. 640–643.

[2] H. Kitano. "Computational systems biology". *Nature* **420**:6912 (2002), p. 206.

[3] U. Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

[4] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. "Stochastic gene expression in a single cell". *Science* **297**:5584 (2002), pp. 1183–1186.

[5] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl. "Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings". *Nature methods* **11**:2 (2014), p. 197.

[6] Z. Fox and B. Munsky. "Stochasticity or Noise in Biochemical Reactions". *arXiv preprint arXiv:1708.09264* (2017).

[7] S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. "Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis". *Nature* **459**:7245 (2009), p. 428.

[8] D. R. Bandura, V. I. Baranov, O. I. Ornatsky *et al.* "Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry". *Analytical chemistry* **81**:16 (2009), pp. 6813–6822.

[9] S. C. Bendall, E. F. Simonds, P. Qiu *et al.* "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum". *Science* **332**:6030 (2011), pp. 687–696.

[10] H. Koeppl, C. Zechner, A. Ganguly, S. Pelet, and M. Peter. "Accounting for extrinsic variability in the estimation of stochastic rate constants". *International Journal of Robust and Nonlinear Control* **22**:10 (2012), pp. 1103–1119.

[11] D. A. McQuarrie. "Stochastic approach to chemical kinetics". *Journal of applied probability* **4**:3 (1967), pp. 413–478.

[12] A. Colman-Lerner, A. Gordon, E. Serra *et al.* "Regulated cell-to-cell variation in a cell-fate decision system". *Nature* **437**:7059 (2005), p. 699.

[13] H. Kitano. "Systems biology: a brief overview". *Science* **295**:5560 (2002), pp. 1662–1664.

[14] L. Weber, W. Raymond, and B. Munsky. "Identification of gene regulation models from single-cell data". *Physical biology* **15**:5 (2018), p. 055001.

[15] N. Le Novere, B. Bornstein, A. Broicher *et al.* "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems". *Nucleic acids research* **34**:suppl_1 (2006), pp. D689–D691.

[16] K. McGoff, S. Mukherjee, N. Pillai *et al.* "Statistical inference for dynamical systems: A review". *Statistics Surveys* **9** (2015), pp. 209–252.

[17] B. Munsky, B. Trinh, and M. Khammash. "Listening to the noise: random fluctuations reveal gene network parameters". *Molecular systems biology* **5**:1 (2009), p. 318.

[18] C. Zechner, J. Ruess, P. Krenn *et al.* "Moment-based inference predicts bimodality in transient gene expression". *Proceedings of the National Academy of Sciences* **109**:21 (2012), pp. 8340–8345.

[19] A. Raue, C. Kreutz, T. Maiwald *et al.* "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood". *Bioinformatics* **25**:15 (2009), pp. 1923–1929.

[20] M. Sunnåker, E. Zamora-Sillero, A. López García de Lomana *et al.* "Topological augmentation to infer hidden processes in biological systems". *Bioinformatics* **30**:2 (2013), pp. 221–227.

[21] M. Sunnåker, E. Zamora-Sillero, R. Dechant *et al.* "Automatic generation of predictive dynamic models reveals nuclear phosphorylation as the key Msn2 control mechanism". *Sci. Signal.* **6**:277 (2013), ra41–ra41.

[22] S. Bandara, J. P. Schlöder, R. Eils, H. G. Bock, and T. Meyer. "Optimal experimental design for parameter estimation of a cell signaling model". *PLoS computational biology* **5**:11 (2009), e1000558.

[23] G. Lillacci and M. Khammash. "Parameter estimation and model selection in computational biology". *PLoS computational biology* **6**:3 (2010), e1000696.

[24] J. Ruess, A. Milias-Argeitis, and J. Lygeros. "Designing experiments to understand the variability in biochemical reaction networks". *Journal of the Royal Society Interface* **10**:88 (2013), p. 20130588.

[25] J. Ruess, F. Parise, A. Milias-Argeitis, M. Khammash, and J. Lygeros. “Iterative experiment design guides the characterization of a light-inducible gene expression circuit”. *Proceedings of the National Academy of Sciences* (2015), p. 201423947.

[26] U. Helmke, K. Huper, and M. Khammash. “Global identifiability of a simple linear model for gene expression analysis”. *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on.* IEEE. 2013, pp. 7149–7154.

[27] A. Gupta and M. Khammash. “Sensitivity analysis for multiscale stochastic reaction networks using hybrid approximations”. *arXiv preprint arXiv:1801.04708* (2018).

[28] A. Gupta, M. Rathinam, and M. Khammash. “Estimation of parameter sensitivities for stochastic reaction networks using tau-leap simulations”. *SIAM Journal on Numerical Analysis* **56**:2 (2018), pp. 1134–1167.

[29] D. T. Gillespie. “Approximate accelerated stochastic simulation of chemically reacting systems”. *The Journal of Chemical Physics* **115**:4 (2001), pp. 1716–1733.

[30] Y. Cao, D. T. Gillespie, and L. R. Petzold. “Adaptive explicit-implicit tau-leaping method with automatic tau selection”. *The Journal of chemical physics* **126**:22 (2007), p. 224101.

[31] B. Munsky and M. Khammash. “The finite state projection algorithm for the solution of the chemical master equation”. *The Journal of chemical physics* **124**:4 (2006), p. 044104.

[32] J. Feigelman, D. Weindl, F. J. Theis, C. Marr, and J. Hasenauer. “LNA++: Linear Noise Approximation with First and Second Order Sensitivities”. *International Conference on Computational Methods in Systems Biology.* Springer. 2018, pp. 300–306.

[33] M. Khammash and J. Lygeros. “Identification and Control of Cell Populations”. *Encyclopedia of Systems and Control* (2015), pp. 547–552.

[34] R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and G. Stolovitzky. *Crowdsourcing network inference: the DREAM predictive signaling network challenge.* 2011.

[35] S. M. Hill, N. K. Nesser, K. Johnson-Camacho *et al.* “Context specificity in causal signaling networks revealed by phosphoprotein profiling”. *Cell systems* **4**:1 (2017), pp. 73–83.

[36] D. Koller, N. Friedman, and F. Bach. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[37] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. “Using Bayesian networks to analyze expression data”. *Journal of computational biology* **7**:3-4 (2000), pp. 601–620.

[38] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger. “Bayesian network approach to cell signaling pathway modeling”. *Sci. STKE* **2002**:148 (2002), pe38–pe38.

[39] J. Pearl and T. S. Verma. “A theory of inferred causation”. *Studies in Logic and the Foundations of Mathematics.* Vol. 134. Elsevier, 1995, pp. 789–811.

[40] J. Pearl. *Causality.* Cambridge university press, 2009.

[41] D. Pe’er, A. Regev, G. Elidan, and N. Friedman. “Inferring subnetworks from perturbed expression profiles”. *Bioinformatics* **17**:suppl_1 (2001), S215–S224.

[42] K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan. “Causal protein-signaling networks derived from multiparameter single-cell data”. *Science* **308**:5721 (2005), pp. 523–529.

[43] X.-K. Lun, V. R. Zanotelli, J. D. Wade *et al.* “Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry”. *Nature biotechnology* **35**:2 (2017), p. 164.

[44] S. Rogers and M. Girolami. “A Bayesian regression approach to the inference of regulatory networks from gene expression data”. *Bioinformatics* **21**:14 (2005), pp. 3131–3137.

[45] K. Sachs, S. Itani, J. Fitzgerald *et al.* “Single timepoint models of dynamic systems”. *Interface focus* **3**:4 (2013), p. 20130019.

[46] S. Itani, M. Ohannessian, K. Sachs, G. P. Nolan, and M. A. Dahleh. “Structure learning in causal cyclic networks”. *Causality: Objectives and Assessment.* 2010, pp. 165–176.

[47] K. Sachs, S. Itani, J. Fitzgerald *et al.* “Learning cyclic signaling pathway structures while minimizing data requirements”. *Biocomputing 2009.* World Scientific, 2009, pp. 63–74.

[48] S. Triantafillou, V. Lagani, C. Heinze-Deml *et al.* “Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells”. *Scientific reports* **7**:1 (2017), p. 12724.

[49] N. Meinshausen, A. Hauser, J. M. Mooij *et al.* “Methods for causal inference from gene perturbation experiments and validation”. *Proceedings of the National Academy of Sciences* **113**:27 (2016), pp. 7361–7368.

[50] S. Bongers and J. M. Mooij. “From Random Differential Equations to Structural Causal Models: the stochastic case”. *arXiv preprint arXiv:1803.08784* (2018).

[51] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT press, 2017.

[52] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect.* Basic Books, 2018.

Part II.

Scientific Contributions

2 Sparse regression based structure learning of stochastic reaction networks from single cell snapshot time series

Anna Klimovskaia, Stefan Ganscha, Manfred Claassen

Originally published in: *PLoS computational biology* 2016 Dec 6;12(12):e1005234.. Adapted here with minor modifications.

Contribution by AK: conceived and implemented the methodology, designed and performed the computational experiments, wrote the manuscript.

2.1. Introduction

Cellular processes are essentially implemented by networks of biochemical reactions. The topology of such networks is typically only partially known, rendering the identification of the correct network from experimental data a key challenge. Despite the importance of this task, only little progress has been made in devising methods to systematically and comprehensively infer topologies of non-trivial chemical reaction networks. In this work, we propose a sparse regression approach tailored to the task of large-scale model selection for chemical reaction networks.

Different model classes have been developed to describe biochemical reaction systems. In order of increasing level of detail these comprise statistical time series models, such as autoregressive models and dynamic Bayesian networks, deterministic ordinary differential equation or stochastic differential equation based kinetic models [1]. The choice of model class depends on prior information for the system of interest and type of experimental data. Single cell technologies furnish further statistical information about component distributions, e.g. variances and covariances, aiding in systems identification [2] and are expected to become increasingly prevalent in routine biological research [3].

Two main computational tasks arise when learning any of these models from data: parameter inference, and structure learning. Parameter inference aims at finding model parameters (e.g. kinetic rate constants). Parameter inference has been performed by sampling from posterior parameter distributions, or global non-convex or convex optimization methods [4]. Structure learning aims at additionally identifying the reaction network topology governing the dynamics of the system components.

Parameter inference becomes increasingly computationally intensive for larger systems with numerous parameters [1]. Structure learning for these systems is an even more daunting task since parameter inference has to be performed for each of the possibly very many different system topologies. Therefore, structure learning is typically confined to comparison of a small, carefully selected set of candidate topologies by means of model selection criteria, such as information criteria (e.g. AIC, BIC) or Bayes Factors [5–7]. However, this approach requires substantial prior knowledge about the studied system in order to identify reasonable candidate models. Systematic approaches to enumerate a subset of sensible topologies have been not reported until recently. These approaches implement greedy strategies that either iteratively reduce the number of reactions of an overcomplete system of reactions or add reactions one at a time to a system with a minimal set of reactions [8]. However, such greedy approaches do not guarantee finding globally optimal topologies for non-convex fitting objectives. Furthermore, exploration of the multitude of local optima in the context of combinatorially many possible topologies becomes computationally prohibitive due to the requirement to explicitly evaluate every considered candidate topology. No global approaches have been reported to perform structure learning by comprehensively evaluating model candidates for stochastic chemical reaction networks.

We propose the *reactionet lasso*, a convex relaxation of the structure learning task. This approach yields a single best sparse reaction set from all possible reactions by translating a recent sparse identification approach for nonlinear dynamic systems [9] to operate on and deal with non-trivial application specific parameter and noise structure for time series snapshot data acquired for stochastic chemical reaction networks.

2.2. Results

2.2.1. Sparse regression for structure learning of stochastic reaction networks

Structure learning by the reactionet lasso takes advantage of the formal link between the chemical reaction model and the observed data that is defined by the Chemical Master Equation. This differential equation system describes the temporal evolution of the abundance distributions of species governed by a stochastic chemical reaction network [10]. The moment generating functions of the Chemical Master Equation give rise to the moment equations, a system of ordinary differential equations for the temporal evolution of the central moments $M_{\mathbf{r}}$ of the abundance distributions (see *Methods*).

$$\dot{M}_{\mathbf{r}} = \Sigma_l k_l F_{\mathbf{r}l}(t; \mathbf{M}), \tag{2.1}$$

with rate constants k_l, time t and set of all central moments of individual species $\mathbf{M}$. For mass action kinetics the terms $F_{\mathbf{r}l}(t; \mathbf{M})$ are polynomials over these moments such as abundance means and variances of individual species. $F_{\mathbf{r}l}(t; \mathbf{M})$ will be referred to as stoichiometric moment functions herein (see also **S1 Text**).

The moment equations constitute the formal link between the time series snapshot data and the rate constants of the underlying chemical reactions. Rate constant estimation for stochastic mass action kinetics reaction networks in this context therefore reduces to parameter estimation for the ordinary differential equation system [2.1] with stoichiometric moment functions determined from the time series data.

Parameter estimation for a mass action kinetics network typically requires the costly integration of the moment equations for every considered parameter configuration. Imputation of the moment gradients by gradient matching procedures (see *Methods*) circumvents these type of evaluations and, in conjunction with the empirical moments, allows for parameter inference by means of a non-negative linear regression task with the least squares estimate $\hat{\mathbf{k}}$ for rate constants $\mathbf{k}$ given by:

$$\hat{\mathbf{k}} = \arg\min_{\mathbf{k}>0} \|\hat{\mathbf{b}} - \hat{\mathbf{A}}\mathbf{k}\|_2^2, \tag{2.2}$$

where the response vector elements b_j corresponds to the vector of empirical gradient estimates for $\dot{M}_j(t)$ from the gradient matching procedure (see *Methods*) and the design matrix entries $\hat{A}_{jl}$ correspond to the estimates of the stoichiometric moment functions $F_{jl}(t;\mathbf{M})$:

$$\mathbf{b} = \begin{bmatrix} \hat{\dot{M}}_{r_1}(t_1) \\ \vdots \\ \hat{\dot{M}}_{r_1}(t_T) \\ \hat{\dot{M}}_{r_2}(t_1) \\ \vdots \\ \hat{\dot{M}}_{r_2}(t_T) \\ \vdots \\ \hat{\dot{M}}_{r_N}(t_1) \\ \vdots \\ \hat{\dot{M}}_{r_N}(t_T) \end{bmatrix}, \mathbf{A} = \begin{bmatrix} \hat{F}_{r_1 1}(t_1) & \hat{F}_{r_1 2}(t_1) & \dots & \hat{F}_{r_1 L}(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{F}_{r_1 1}(t_T) & \hat{F}_{r_1 2}(t_T) & \dots & \hat{F}_{r_1 L}(t_T) \\ \hat{F}_{r_2 1}(t_1) & \hat{F}_{r_2 2}(t_1) & \dots & \hat{F}_{r_2 L}(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{F}_{r_2 1}(t_T) & \hat{F}_{r_2 2}(t_T) & \dots & \hat{F}_{r_2 L}(t_T) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{F}_{r_N 1}(t_1) & \hat{F}_{r_N 2}(t_1) & \dots & \hat{F}_{r_N L}(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{F}_{r_N 1}(t_T) & \hat{F}_{r_N 2}(t_T) & \dots & \hat{F}_{r_N L}(t_T) \end{bmatrix}.$$

This linear regression formulation has been applied for parameter inference of deterministic chemical reaction models [6, 11].

Model selection across small sets of model variants has previously been performed with information criteria [11] or model averaging [6]. The Lasso constitutes another approach for efficient and comprehensive model selection in linear regression models [12]. It introduces an L1 norm ($\|\cdot\|_1$) regularization on the parameters $\mathbf{k}$ to promote the identification of sparse solutions, i.e. solutions with many zero-valued parameter estimates.

$$\hat{\mathbf{k}} = \arg\min_{\mathbf{k}>0} \|\hat{\mathbf{b}} - \hat{\mathbf{A}}\mathbf{k}\|_2^2 + \lambda\|\mathbf{k}\|_1, \tag{2.3}$$

Various extensions of the Lasso method were introduced in literature to improve its shrinkage properties in the presence or absence of heteroscedasticity [13].

While the Lasso has been used in recent reports to identify general nonlinear dynamical systems [9] or to select the mechanism types (mass action or Hill kinetics) of a fixed reaction set defined by the deterministic Repressilator comprising six components [14], it still remains to adapt the regression model and regularization concepts to enable more comprehensive model selection for realistic reaction systems that exhibit stochasticity and larger amount of components/reactions. The next sections will delineate in detail the challenges and solutions implemented in the reactionet lasso to achieve this goal.

2.2.2. Reactionet lasso

This section introduces the reactionet lasso (**Figure 2.1**), a computational method for learning the structure of chemical reaction networks. The overarching strategy of this procedure consists of (1) enumerating all (or at least a significant fraction of reasonable) conceivable unary/binary reactions between the components of a reaction system of interest and (2) applying an appropriate stepwise sparse regression approach to select the sparse subset of reactions underlying the observed dynamics in the snapshot time series data.

The following properties of such structure learning instances preclude the application of conventional least squares based approaches for parameter estimation and selection: (1) noise and heteroscedasticity of the observed response $\hat{\mathbf{b}}$ (empirical moment gradient estimates) as well as in the observed design matrix $\hat{\mathbf{A}}$ (stoichiometric moment function evaluations) and (2) different scales of individual parameters k_i (rate constants) resulting from the occurrence of large a spectrum of fast and slow reactions. The reactionet lasso addresses each of these challenges in as delineated in the following.

The intrinsic variability of stochastic chemical kinetics result induce variability of the

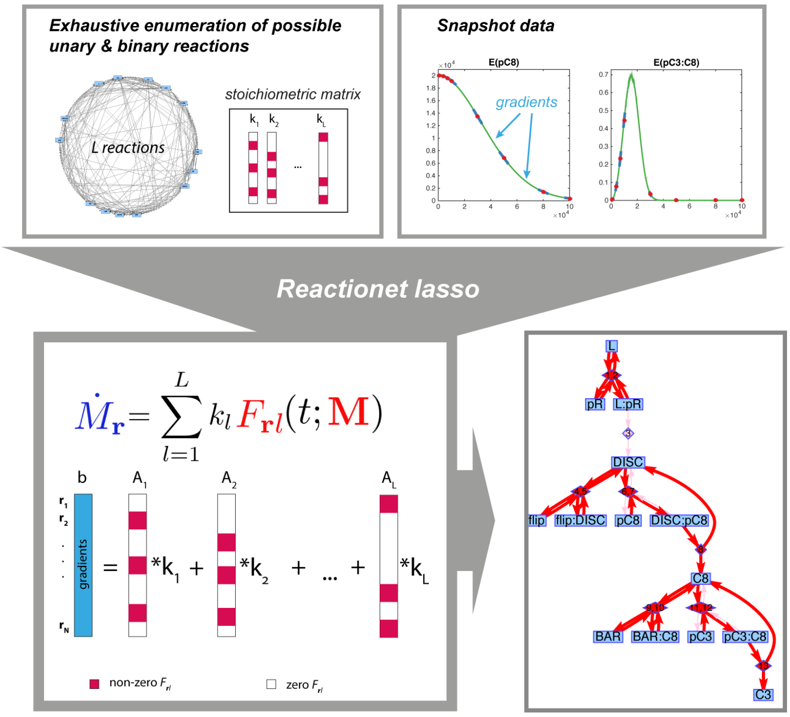

Exhaustive enumeration of possible unary & binary reactions
L reactions
stoichiometric matrix
k1 k2 ... kL
Snapshot data
E(pC8)
E(pC3:C8)
gradients
Reactionet lasso
$\dot{M}_{\mathbf{r}} = \sum_{l=1}^{L} k_l F_{\mathbf{r}l}(t; \mathbf{M})$
b A1 A2 AL
r1 r2 . . . rN
gradients
= *k1 + *k2 + ... + *kL
non-zero F_{rl}
zero F_{rl}
L
pR L:pR
DISC
flip flip:DISC pC8 DISC:pC8
C8
BAR BAR:C8 pC3 pC3:C8
C3

Figure 2.1.: **Schematic representation of *reactionet lasso* procedure.** For details see section Methods.

empirical estimates of moments and their gradients. Therefore the observed response vector as well as the stoichiometric moment functions in the design matrix are expected to deviate from the true latent correspondents. We capture this by defining $\mathbf{b} = \hat{\mathbf{b}} + \epsilon_b$ and $\mathbf{A} = \hat{\mathbf{A}} + \epsilon_A$ to be the true latent moment gradients and stoichiometric moment functions, and ϵ_A and ϵ_b to be their respective intrinsic variability induced deviations from the estimated/observed quantities. If we knew the true values of the latent variables, finding the rate constants $\mathbf{k}$ would translate to solving the following equation:

$$\mathbf{b} = \mathbf{A}\mathbf{k}. \tag{2.4}$$

By substituting the variables in equation 2.4 with the definitions for our empirical estimates of the latent variables we obtain:

$$\hat{\mathbf{b}} = \hat{\mathbf{A}}\mathbf{k}. \tag{2.5}$$

with $\epsilon := \epsilon_A \mathbf{k} - \epsilon_b$.

Equation 2.5 seems to motivate a straightforward optimization strategy to compute a maximum likelihood parameter estimate given the observations for moment gradients and stoichiometric moment functions (e.g. least squares for independent and normally distributed residuals ϵ). However, it becomes apparent that this strategy is not valid due to the residual ϵ being a function of the parameters $\mathbf{k}$ (by virtue of the noise in the observed design matrix).

The reactionet lasso implements a stepwise strategy to address this dependency. The first step (Step 1) is a Feasible Generalized Least Squares (FG) estimate. It comprises the estimation of the variances of the residuals ϵ_b and ϵ_A via bootstrapping of the gradient estimates and stoichiometric moment functions on the basis the single-cell data. A preliminary least squares fit is then performed to achieve an estimate $\mathbf{k}^{LS}$ for equation 2.5. This estimate is expected to approximate the order of magnitude of the individual rate constants. In conjunction with the estimates of the variances of the residuals ϵ_b and ϵ_A, we use $\mathbf{k}^{LS}$ to achieve an estimate of the component-wise variance $\Sigma_\epsilon = \operatorname{diag}\{\sigma^2_{\epsilon_1}, \dots, \sigma^2_{\epsilon_R}\}$ of the residuals ϵ. To achieve this estimate we use only first order moments (means), as they are less subjected to noise in the design matrix and provide a more robust estimate of the covariance matrix Σ_ϵ. This estimate will allow us to operate with the rescaled observed response vector $\hat{\mathbf{b}}^S = \Sigma_\epsilon^{-1/2}\hat{\mathbf{b}}$ and design matrix $\hat{\mathbf{A}}^S = \Sigma_\epsilon^{-1/2}\hat{\mathbf{A}}$ to adjust for heteroscedasticity and enable effective linear regression [15].

$$\hat{\mathbf{k}}^{FG} = \arg\min_{\mathbf{k}>0} \|\hat{\mathbf{b}}^S - \hat{\mathbf{A}}^S\mathbf{k}\|_2^2, \tag{2.6}$$

The subsequent steps aim at addressing the second challenge introduced above, i.e. the different scales of individual parameters k_i, which render conventional sparse regression approaches (such as the Lasso) suboptimal due to the uniform penalization strength of the L1 norm $\|.\|_1$ across all components k_i of the parameter vector k. The adaptive Lasso [16] constitutes an alternative to the conventional Lasso. It defines a regularization penalty that is scaled component-wise by the expected order of magnitude $\hat{k_i}$ of the respective component i.

In Step 2 of the reactionet lasso, we apply a combination of the adaptive and relaxed Lasso, stability selection based prioritization of reactions and an additional stepwise backward regression to achieve the final set of reported reactions. We use the parameter estimates from Step 1 (obtained with Moore–Penrose pseudoinverse matrix), i.e. $\tilde{\mathbf{k}} = \mathbf{k}^{\hat{FG}}$, in order to adapt the regularization penalty.

To improve shrinkage, the adaptive Lasso is followed by a relaxed Lasso [17] that recomputes optimal parameter estimates with respect to the objective specified in equation 2.6, while only considering the set $\Phi = \{l : k_l^{FG} \neq 0\}$ of parameters that were not set to zero in Step 1, for which the optimal solution is

$$\hat{\mathbf{k}}^{ARL} = \arg\min_{\mathbf{k}>0} \|\hat{\mathbf{b}}^S - \hat{\mathbf{A}}^{S,\Phi}\mathbf{k}\|_2^2 + \lambda\Sigma_i|k_i/\tilde{k}_i|, \tag{2.7}$$

where $\hat{\mathbf{A}}^{S,\Phi}$ contains only that columns of $\hat{\mathbf{A}}^S$, which are in a set Φ.

The adaptive relaxed Lasso solution has been computed by optimizing the respective Alternating Direction Method of Multipliers (ADMM) formulations [18]. The adaptive relaxed Lasso is performed with five fold cross validation. We used stability selection to prioritize reactions according to their frequency of being selected across all cross validation folds [19]. Bayesian information criterion (BIC) was used as selection criterion (**S2 Text**).

In summary, the reactionet lasso procedure constitutes a stepwise sparse regression approach that addresses the parameter-dependent noise and heteroscedasticity in the response and design matrix for structure learning of stochastic chemical reaction systems. See also **Figure 2.1** for a schematic overview of its steps. Software implementing the reactionet lasso can be found at `http://www.imsb.ethz.ch/research/claassen/Software/reactionet_lasso.html`.

2.2.3. Ab initio structure learning of chemical reaction networks

We first consider an extreme and yet conceptually simple scenario where we aim at learning the structure of a reaction network without any prior knowledge about the underlying reactions. While this scenario rarely occurs in a real world application because typically some prior knowledge of relevant reactions is available, we first investigate this scenario to demonstrate the structure learning capabilities of the reactionet lasso.

We study two systems varying in number of components and reactions: (1) the enzymatic reaction system with four components and three reactions, (2) the receptor subunit of a recently reported kinetic model of TRAIL induced apoptosis with fourteen components and thirteen reactions, which can be combined in a total of 2275 possible unary or binary reactions, giving a total of more than 10^{600} possible reaction network candidates. For these systems we simulated 5 replicates each with either 10^3, 10^4 or 10^5

single cell trajectories with the stochastic simulation algorithm [20]. We then generated snapshot time series datasets from the single cell trajectories by defining pools of cells at selected sets of 7, 13, or 28 time points. Moment gradients were estimated either with the smoothing procedure, cubic splines or the finite difference scheme (see *Methods*).

The reactionet lasso achieves structure learning of chemical reaction networks via a two step sparse regression formulation that (1) specifically accounts for heteroscedasticity in the response vector and the design matrix of the regression instances and (2) assumes a regularizer that encourages sparse reaction sets by suppressing compensatory reaction sets with small rate constants (**Figure 2.1**). The first step aims at accounting for heteroscedasticity and, most importantly at reducing the number of reaction candidates for the second step that both capture the empirical moment gradients and select for correct reactions (**Figure 2.2**). The following results are based on moment equations for all moments up to order two, i.e. means, variances and covariances. Following Step 1 of the reactionet lasso, we achieve a substantial reduction to less than 100 candidate reactions that, regardless of the moment gradient estimation technique, retains at least ten of the thirteen true reactions (**Figure 2.1A**). The vast majority of the empirical moment gradients are well fit by the set of candidate reactions. The few moment gradients that are suboptimally captured correspond to higher order moments such as variances or covariances whose highly dynamic behavior precluded accurate gradient estimation by either the finite difference or spline fit. (**Figure 2.2B**). Step 2 of the reactionet lasso procedure uses a relaxed adaptive Lasso estimator to estimate the rate constants of a sparse set of candidate reactions following from Step 1. The method recovers ten out of thirteen reactions correctly with one false positive reaction when assuming no prior knowledge and selecting a suitable model with BIC (**Supplementary Figure A.1A**). Similar performance is achieved for the enzymatic reaction network (**Supplementary Figure A.1B**). These results demonstrate that the stepwise sparse regression strategy of Step 2 completes the structure learning task from the candidate reactions supplied by Step 1 with great sensitivity and specificity. In summary, the reactionet lasso is able to ab initio reconstruct the reaction network structure of typically-sized signaling cascades such as the fourteen component receptor subunit of TRAIL induced apoptosis [21].

We further evaluated the impact of different gradient estimation approaches on structure learning performance (**Supplementary Figures A.2-A.4**). For benchmarking purposes we used the smoothed empirical moment gradient estimate as a ground truth which is not available in a real time series snapshot setting. According to these consid-

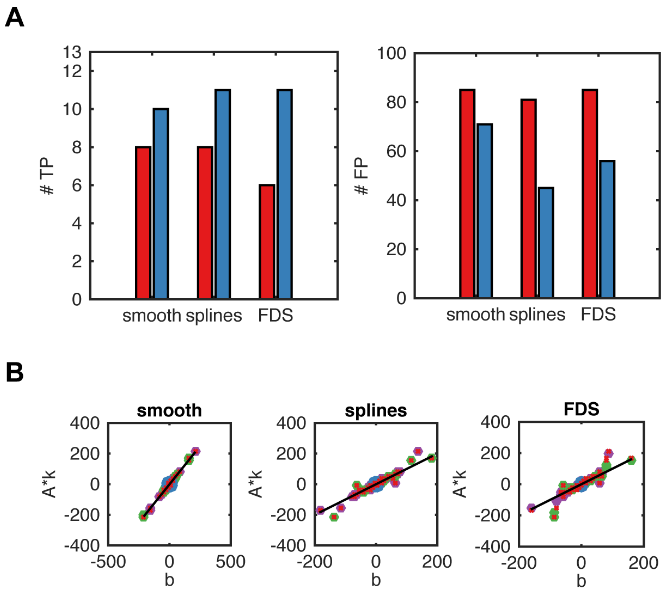


Figure 2.2.: **Performance assessment of the first *reactionet lasso* step for** 10^5 **single cell trajectories of the apoptotic receptor subunit.** (A) Enrichment and depletion of true and respectively false positive reactions for the reaction rate estimates $\hat{\mathbf{k}}^{LS}$ (red) and $\hat{\mathbf{k}}^{FG}$ (blue). Results are reported for gradient estimation procedures smooth, FDS, splines (see main text for details). (B) Comparison of response (empirical moment gradients) and prediction with feasible generalized least square estimate for moments of different order: means (blue), variances (green), covariances (yellow) and prediction with true rate constants for all moments (red crosses).

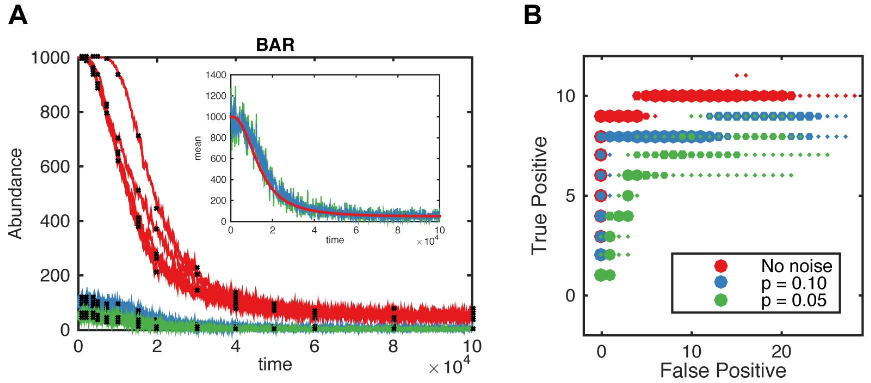


Figure 2.3.: **Influence of measurement noise on structure learning capacity.** (A) Example of several single cell trajectories of one of the species (BAR) in apoptotic receptor subunit: without measurement noise (red), with measurement noise according to the binomial model with probability of success $p = 0.1$ (blue), 0.05 (green). Comparison of reconstructed means for known p between different noise levels shows how empirical moments are affected by measurement noise. Black dots represent snapshot measurements used for the inference procedure. (B) Overlay of five regularization paths in terms of true/false positive tradeoff for different measurement noise levels as indicated in the legend in terms of binomial capture efficiency. Structure learning performance for 10^5 single cell trajectories and thirteen time points of the apoptotic receptor subunit. Empirical moment gradients estimated with splines.

erations, the cubic spline estimator achieves almost optimal performance for thirteen or more time points, whereas FDS is consistently inferior. These results indicate that the cubic spline estimator provides the most favorable structure learning performance for for empirical moment gradients.

We evaluated how measurement noise affects the ability of reactionet lasso to learn the reaction network structure. We assume a binomial measurement noise model that reflects the incomplete capture efficiency inherent to all single cell technologies (see **Methods, S3 Text**). While structure learning performance is reduced with increasing levels of measurement noise, the reactionet lasso still recovers more than 50% of the reactions for the apoptotic receptor subunit at levels reported for single cell sequencing and mass cytometry approaches (**Figure 2.3, S5 Figure**).

To assess the relative importance of the amount of available data, we varied the amount of time points and single cell recordings used at each time point. Interestingly, we found that the inclusion of additional measurement time points did not improve structure learning performance. However, the tradeoff between true and false positive

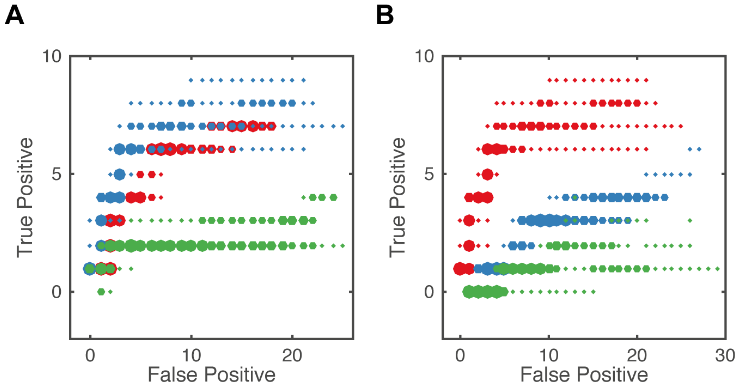


Figure 2.4.: **Overlay of five regularization paths in terms of true/false positive tradeoff over different data availability situations.** Results for *reactionet lasso* application to apoptotic receptor subunit ($p = 0.05$). Empirical moment gradients estimated with cubic splines. (A) 10^5 single cell trajectories evaluated at different amount of time points: 28 (red), 13 (blue), 7 (green). (B) Different number of single cell trajectories: 10^5 (red), 10^4 (blue), 10^3 (green) evaluated at thirteen time points.

reaction discoveries worsened considerably with fewer time points (**Figure 2.4A**). While we found that decreasing the amount of single cell measurements per time point did result in noticeable performance losses, this situation does not constitute a limitation for flow cytometry techniques, that easily are able to generate millions of single cell snapshots (**Figure 2.4B**). Cell count related performance losses can be associated with higher absolute variability and therefore reduced accuracy of empirical moment estimates (**Supplementary Figure A.6**). We conclude that careful selection of amount of single cell measurements and number as well as position of time points (**Supplementary Figure A.7**) translate to accurate interpolation and subsequent gradient fitting, thereby leading to good structure learning performance of the reactionet lasso.

We further investigated the impact of including different moment orders for structure learning. As expected, precisely estimated higher-order moments contain a substantial amount of information and therefore enhance the structure learning capability accordingly (**Figure 2.5A**). However, although this relationship still holds for medium levels of measurement noise (capture efficiency $p = 0.1$), (**Figure 2.5B**), the inclusion of second order moments becomes misleading for high levels of measurement noise (capture efficiency $p = 0.05$, **Figure 2.5C**). This observation is likely caused by the limited ability to accurately estimate higher order moments for high levels of measurement noise. However, the performance of the reactionet lasso assuming stochastic kinetics modeled

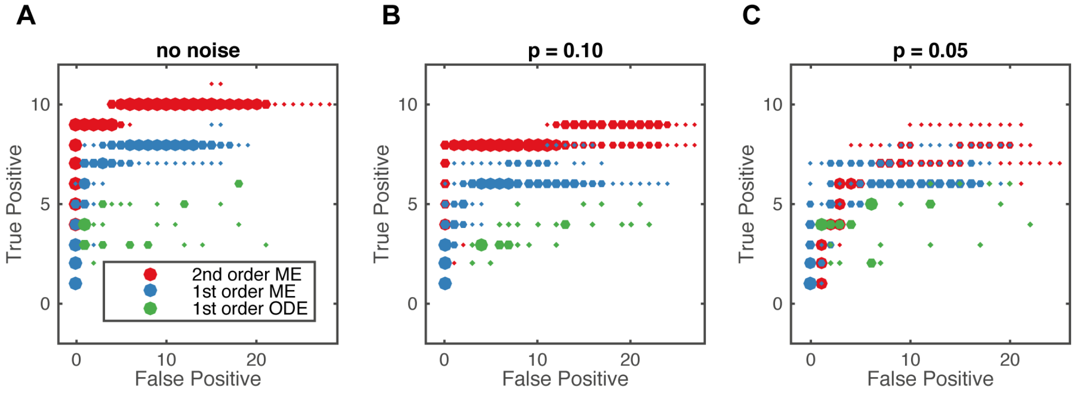


Figure 2.5.: **Overlay of five regularization paths in terms of true/false positive tradeoff for different moment orders of gradients considered for structure learning:** up to 2nd order (red) Moment Equations, up to 1st order (blue) Moment Equations and only 1st order moments for deterministic mean ODE model (green). Structure learning performance for 10^5 single cell trajectories and thirteen time points of the apoptotic receptor subunit. Empirical moment gradients estimated with splines. Results represented for different levels of measurement noise: (A) no noise; (B) $p = 0.1$; (C) $p = 0.05$.

with moment equations (higher order ME) is consistently better than assuming a deterministic kinetics modeled with mean based ordinary equations (1st order ODE). This observation demonstrates that the incorporation of higher order moment information induced by the chemical kinetics and accessible by means of single cell measurements allows for significantly improved structure learning capacity.

In summary, the benchmarks above strongly advocate for the use of an experimental setup that allows for sufficiently dense sampling across time to ensure accurate empirical moment gradient estimates, as well as single cell technology, such as flow/mass cytometry, which provide 10^4 or more single cell measurements, for the accurate estimation of higher order moments. In these situations the reactionet lasso is capable of ab initio recovery of almost the complete reaction network structure with more than a dozen components.

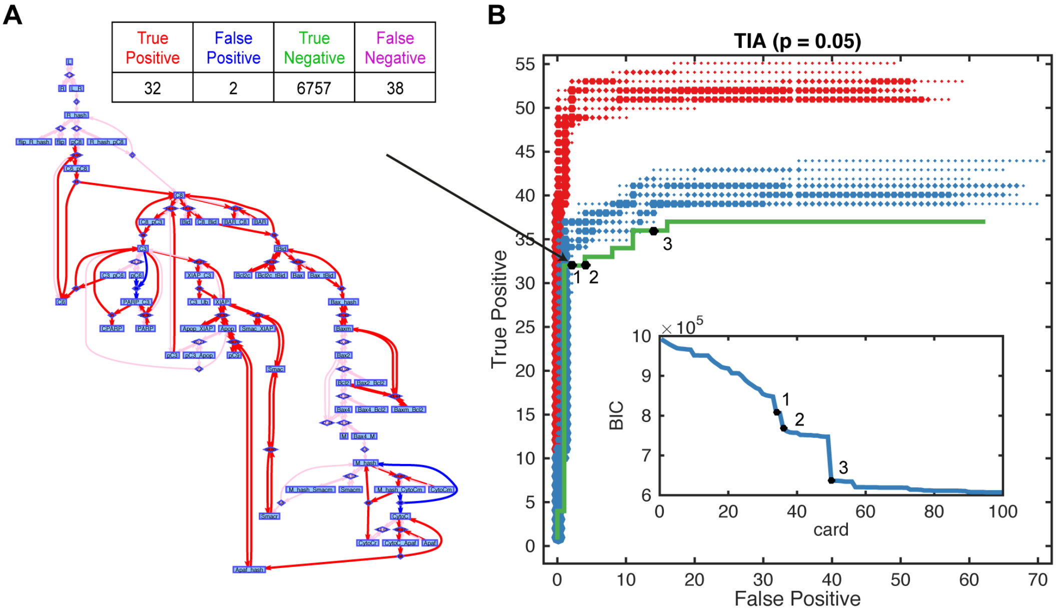
2.2.4. Structure learning of large chemical reaction networks with prior knowledge

We now consider a scenario where we aim at learning the structure of a large reaction network with partial knowledge about the underlying reactions. For this situation we demonstrate how reactionet lasso is capable of recovering a sizable amount of the unknown reactions, for a reaction network as large as the 70 reaction TRAIL induced apoptosis cascade [21].

Structure learning tasks for chemical reaction networks typically aim at complementing already available partial knowledge on reaction sets. We assessed the ability of the reactionet lasso to complement a set of known reactions for the 70 reaction TRAIL induced apoptosis cascade. Specifically, we defined six modules for this cascade following [21], and assumed a limited set of 22 reaction candidates connecting these modules (**Supplementary Figure A.8**) and 33 uniformly distributed time points (S1 Dataset). For step 1 of the reactionet lasso all possible unary and binary reactions between components within modules and the module connecting reactions serve as candidate reactions for structure learning, totaling 6828 reactions.

In the absence of ground truth it is difficult to identify a regularization strength that achieves a desirable tradeoff between true and false positive reaction discoveries. We evaluated the BIC and report solutions that map to large initial improvements of BIC [22]. Structure learning without prior knowledge on the considered set of reactions achieves 32 true positive at the cost of 2 false positive reactions (10^5 single cell trajectories, 33 time points, capture efficiency 0.05, **Figure 2.6A**). Prior knowledge on a specific reaction was encoded by a positivity constraint on the corresponding reaction rate during all regression steps of the reactionet lasso. We considered different prior knowledge settings: (1) 10% or (2) 50% randomly chosen reactions considered to be known. Settings (1) and (2) were each evaluated using ten different subsets. For 10% known reactions almost 40 (including 7 known) out of 70 reactions are correctly recovered with five or less false positive discoveries (**Figure 2.6B**). For 50% known reactions the total number of true positive reactions is beyond 50 (including 35 known). The performance doesn't depend significantly on the choice of prior reactions. The reactionet lasso enables discovery up to dozens of novel reactions at the cost of few false positive reactions for a large signaling cascade comprising 70 reactions.

While published structure learning approaches are only available for problem instances



True Positive	False Positive	True Negative	False Negative
32	2	6757	38

Figure 2.6.: **Structure learning with prior knowledge for the 70 component TRAIL induced apoptosis cascade.** Structure learning performance for 10^5 single cell trajectories and 33 time points and capture efficiency $p = 0.05$. Empirical moment gradients estimated with splines. (A) Example of a recovered graph for the setting above and no prior knowledge. True positive reactions are red, false positive reactions are blue, false negative reactions are pink. (B) Regularization paths in terms of true/false positive tradeoff (including prior knowledge reactions, see **Results** for details) for different prior knowledge situations. Following prior knowledge situations are depicted: no prior knowledge (green). Additional prior knowledge situations comprise ten instances of 10%(blue)/ 50%(red) randomly chosen known reactions. Diameter of dots and color code indicate frequency of solutions with a specific true/false positive tradeoff. Black dots represent solutions coinciding with large improvements of BIC.

of sizes hundreds of orders of magnitude smaller, we compared these results to more simple variants of the reactionet lasso procedure, either exhibiting inferior accuracy or exceedingly high computational complexity (**Supplementary Figure A.9**).

The above results demonstrate how single cell snapshot time series data and the reactionet lasso can be used to complement prior mechanistic knowledge by a sizable set of candidate reactions that is highly enriched for true positive discoveries, and do so for systems and structure learning tasks of unprecedented size [23, 24].

2.3. Discussion

In this work we introduce the reactionet lasso for comprehensive structure learning of stochastic chemical reaction networks.

Chemical reaction networks constitute a highly detailed and mechanistic description for biological processes and are qualitatively different from other popular network models in biology. These comprise probabilistic graphical models seeking to discover statistical dependencies between measured system components. These approaches range from simple correlation [25] or regression analysis [26], to Bayesian networks [27] or more structured and robust module networks [28] and extensions thereof [29]. In contrast to chemical reaction networks, each of these model classes allows for detection of statistical dependencies without further elucidation of causality relationships and the possibly intricate dependency inducing biochemical mechanism. Physical interaction networks get closer to this goal and complement the information of reaction networks by summarizing measurements of static protein interactions [30].

By virtue of formulating the task of structure learning of chemical reaction networks as a sequence of convex optimization problems, this procedure is able to assess an unprecedented number of potential network topologies without need for explicit enumeration [23, 24]. We demonstrate the utility of the method for ab initio structure learning of whole signaling cascades such as the apoptotic receptor subunit. The reactionet lasso originally integrates a moment based description of stochastic reaction networks with sparse regression approaches via a gradient matching to achieve an efficient and scalable structure learning procedure, overcoming the limitations of available methods for structure learning which either explicitly enumerate a small set of models or greedily search for locally optimal topologies [8, 31]. Recent generic sparse regression approaches for identification of general nonlinear dynamical systems are in principle applicable for structure learning of biological reaction networks [9]. However, these approaches, in contrast to the reactionet lasso, do not take into account their (1) foundation in the Chemical Master Equation, (2) heteroscedastic and parameter dependent noise structure, as well as (3) parameter ranges varying across many scales, therefore failing to achieve competitive structure learning performance (**Supplementary Figure A.9**).

The challenging structure learning task crucially depends on sensible experimental design yielding informative data. A central design choice concerns the selection of time points recording the relevant dynamical changes of the process of interest. These are typically chosen from prior knowledge or preliminary dense snapshot time series experiments with a cheap readout, such as population based instead of single cell measurements. Another important experimental parameter concerns the number of single cell snapshots. Our benchmarks advocate for having at least thousands of snapshots per time point. Flow cytometry experiments easily achieve snapshot counts in the order of 10^5. For single cell transcriptomics experiments it seems advisable to resort to novel droplet based techniques achieving $> 10^4$ single cell snapshots per experiment [32, 33].

Structure learning performance of the reactionet lasso depends on the accuracy of the gradient estimates from the time series snapshots. We find that estimates based on gradients obtained by rather simple approaches such as finite difference approximations or spline curve fitting achieve competitive performance. Improvements are conceivable by resorting to other techniques specifically designed for gradient estimation in differential equation systems [34, 35]. These approaches jointly fit parameters of the curve fitting procedure and the differential equation system. The success of this strategy relies on considering a problem instance where the differentiation equation systems strongly constrains the state space. However, the problem instances we consider for systematic structure learning with the reactionet lasso assume a differential equation system defined by the moment equation for all possible unary and binary reactions. Such a system will by definition impose little constraints on the state space. Afore gradient matching approaches would therefore have to be adapted to avoid expected parameter overfitting resulting from their application to problem instances with such an expressive differential equation system.

For this proof of principle study, we consider single time series experiments. Reactionet lasso analysis easily accommodates multiple replicates or perturbation experiments such as dose responses. Specifically, condition specific response vectors b_k and design matrices

A_k for each condition k are utilized to construct a problem instance by concatenation. For this problem instance reactionet lasso can be applied as described (see also **S4 Text**). Additional experiments are expected to enhance structure learning performance. Indeed, we observe that this is the case for incorporating additional replicate time series experiments (**Supplementary Figure A.10**).

The reactionet lasso is able to recover a significant proportion of missing reactions in various settings. However, integration of the moment equations for the component means assuming this set is not always able to recover the observed temporal dynamics of the system. This situation arises for instance when a single pivotal true reaction is missed and therefore precludes the correct reconstruction of downstream component dynamics. We frequently encounter this situation in the ab initio structure learning scenario (**Supplementary Figure A.11**). This scenario though constitutes an artificial setting that we only report for a proof of concept of the reactionet lasso. Real world applications comprise prior knowledge about true reactions, typically comprising specifically those pivotal reactions. It turns out that we achieve good reconstruction of integrated trajectories for the structure learning settings assuming prior knowledge (**Supplementary Figure A.11**).

Until now we consider reaction systems which obey mass action kinetics. Systems of this kind can be easily translated into a series of moment equations which depend linearly upon the reaction rates However, systems with non-mass action kinetics, such as Michaelis-Menten, can still be addressed with the reactionet lasso. While appropriate moment closure approximations for certain rational rate law kinetics preserve convexity of the reactionet lasso objective [36], generally such kinetics might yield non-convex optimization problems that would have to be dealt with using appropriate optimization techniques.

The reactionet lasso generates a single point estimate for the optimal, sparse reaction network that might neglect other reasonable candidate network structures. Thus, it will be interesting to perform further in-depth analysis of the resulting network structures, for instance with Markov Chain Monte Carlo sampling techniques.

For our study we assume that all relevant molecular components can be measured. Many biological applications, however, do not allow monitoring all relevant components, as for instance antibodies might only be available for a subset of components of a signaling cascade. While the aim of our study was to demonstrate proof of concept for large scale structure learning of chemical reaction networks, it will be possible to account for missing measured components by either augmenting the model by introducing latent variables or 'lumping' them into more complex non-mass action reaction mechanisms [37].

The reactionet lasso can be applied in its current form to systems where a significant proportion of relevant components can be measured. Considering the steady advance of single-cell technologies, we expect an increasing number of cellular signaling and metabolic processes to be assayed at single-cell resolution. While mass cytometry approaches allow for measurement of more than 30 protein components, sufficient e.g. to substantially map out the T cell receptor, epidermis growth factor and apoptosis signaling cascades, single-cell RNA sequencing opens the prospect of achieving genome-wide transcriptomic snapshots of single cells. Thus we anticipate a surge of relevant data in the near future for which the reactionet lasso can straightforwardly be applied for systamatic and comprehenisve structure learning of the underlying reaction networks, with direct implications for systems biology and health by providing quantitative and predictive models for scientific insight and rational intervention design.

2.4. Methods

2.4.1. Experimental setting: single cell time series snapshot data

We assume time series data with single cell resolved population snapshots obtained at discrete time points. We denote by C the number of cells measured per experiment, T the number of time points at which measurements were performed and N the number of components (e.g. proteins) measured in each cell. For each measurement time point $t = 1, \dots, T$ for a cell $c = 1, \dots, C$ we denote a vector of measured N protein abundances $\mathbf{x}_t^c = \{x_{t,1}^c, \dots, x_{t,N}^c\}$. Therefore at each time point t vectors $x_{t,1}^c, \dots, x_{t,N}^c$ represent a sample from a high-dimensional distribution, which evolves according to the Chemical Master Equation.

2.4.2. Moment equations for the Chemical Master Equation

We assume a biochemical reaction network of N different chemical species with abundances $X_1, \dots, X_N$ involved in L reactions. Each reaction l is characterized by stoichiometry vector s_l and propensity function $a_l(\mathbf{x}; k_l)$ with $\mathbf{x}$ representing the collection of species abundances (system state) and k_l the reaction rate. In our work we consider systems described by mass action kinetics, resulting propensities $a_l(\mathbf{x}; k_l) = k_l * g_l(\mathbf{x})$, where $g_l(\mathbf{x})$ is a known function of the system's state. The state of the system evolves probabilistically according to the possible reactions, with probability $P(\mathbf{x}, t)$ of occupying state $\mathbf{x}$ at time t. The probabilistic evolution of the system's state is described by the Chemical Master Equation:

$$\frac{P(\mathbf{x}, t)}{dt} = \Sigma_{l=1}^{L} P(\mathbf{x} - \mathbf{s}_l, t) a_l(\mathbf{x} - \mathbf{s}_l) - P(\mathbf{x}, t) a_l(\mathbf{x}). \tag{2.8}$$

We denote by $M_{\mathbf{r}} = M_{r_1,...,r_N} = E(X_1 - EX_1)^{r_1} \ldots (X_N - EX_N)^{r_N}$ the central moment of order $\mathbf{r} = (r_1, ..., r_N)$. The moment generating function of the probability distribution $P(\mathbf{x}, t)$ can be used for the derivation of moment equations [38]. Assuming mass action kinetics, we obtain Eq. 2.1 for the time evolution of a central moment (see also **S1 Text**).

2.4.3. Gradient matching for parameter estimation of ordinary differential equation systems

Gradient matching approaches avoid costly integration by instead interpolating the discrete snapshot time series data and estimating the empirical moment gradients $M_{\mathbf{r}}(t)$, rendering the initial ODE system an algebraic equation system with the parameters as unknowns. This formulation further eliminates the need for moment closure, in contrast to integration based techniques. Previously, gradients have been estimated with spline interpolators [34, 39, 40], Gaussian processes [35, 41] or finite difference approximations [6]. Parameter estimation has been performed by least squares minimization [34, 39, 40] or by approximation of the parameter posterior [6, 35, 41]. While deterministic chemical reaction networks frequently served as application settings for gradient matching schemes, only little attention has been paid to networks with stochastic dynamics [6, 11].

We used and compared cubic spline interpolators (spline) and finite difference approximations (FDS) to estimate empirical moment gradients for the $M_{\mathbf{r}}(t)$ of the moment equations. As a ground truth estimate for simulated data, we use a smoothed finite difference approximation of the single cell trajectories at the evaluation point of interest (smooth). Gradient estimates are obtained via a smoothing procedure that relies on a sliding window estimate of finite differences on the simulated trajectories using the smoothing function smooth in Matlab.

2.4.4. Moment estimation for noisy single cell data

Single cell data such as obtained from flow/mass cytometry and single cell sequencing exhibit measurement noise. These technologies each detect a random fraction of the total molecular content of every individual cell. This relationship between the measurement signal and cellular analyte abundance has been frequently modeled by a binomial distribution $Bi(X, p)$ whose success probability p corresponds to the capture efficiency for the analyte present at amount X [42, 43]. We have devised an estimator to subtract the misleading measurement noise component to provide the reactionet lasso with the appropriate noise-correct empirical moment estimates for structure learning.

We assume that measurement noise can be represented by the following binomial model. Let X represent the true abundance of one species at a given time point. Let X^{obs} be the corresponding measured signal, such that $X^{obs} \sim Bi(X, p)$, where p is the capture efficiency. The binomial noise model allows for specifying the following analytical relationships between the first and second order moments of X and X^{obs}:

$$E[X^{obs}] = pE[X], \tag{2.9}$$

$$E[(X^{obs})^2] = p(1-p)E[X] + p^2E[X^2], \tag{2.10}$$

$$Var[X^{obs}] = p(1-p)E[X] + p^2Var[X], \tag{2.11}$$

$$\text{cov}(X_1^{obs}, X_2^{obs}) = p^2\text{cov}(X_1, X_2), \tag{2.12}$$

For a derivation see **S3 Text** . We assume that the capture efficiency p of the single cell instrument is known [42, 43] and estimate the empirical moments of X on the basis of the empirical moments of X^{obs} by solving the above equations for the respective moment of X. The resulting moment estimates are then used in the regression procedure described above to perform structure learning.

2.5. Acknowledgments

We acknowledge Justin Feigelman, Will Macnair, Eirini Arvaniti and Dimitris Christodoulou for helpful discussions and feedback on the manuscript.

References

[1] A. Raue, M. Schilling, J. Bachmann *et al.* "Lessons learned from quantitative dynamical modeling in systems biology". *PLoS One* **8**:9 (2013), e74335.

[2] C. Zechner, J. Ruess, P. Krenn *et al.* "Moment-based inference predicts bimodality in transient gene expression". *Proceedings of the National Academy of Sciences* **109**:21 (2012), pp. 8340–8345.

[3] L. de Vargas Roditi and M. Claassen. "Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics". *Curr. Opin. Biotechnol.* **34C** (2014), pp. 9–15.

[4] C. G. Moles, P. Mendes, and J. R. Banga. "Parameter estimation in biochemical pathways: a comparison of global optimization methods". *Genome Res.* **13**:11 (2003), pp. 2467–2474.

[5] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling. "Ensemble modeling for analysis of cell signaling dynamics". *Nat. Biotechnol.* **25**:9 (2007), pp. 1001–1006.

[6] C. J. Oates, F. Dondelinger, N. Bayani *et al.* "Causal network inference using biochemical kinetics". *Bioinformatics* **30**:17 (2014), pp. i468–74.

[7] P. Kirk, T. Thorne, and M. Stumpf. "Model selection in systems and synthetic biology". *Curr. Opin. Biotechnol.* (2013).

[8] M. Sunnåker, E. Zamora-Sillero, R. Dechant *et al.* "Automatic Generation of Predictive Dynamic Models Reveals Nuclear Phosphorylation as the Key Msn2 Control Mechanism". *Sci. Signal.* **6**:277 (2013), ra41.

[9] S. L. Brunton, J. L. Proctor, and J. N. Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". en. *Proc. Natl. Acad. Sci. U. S. A.* **113**:15 (2016), pp. 3932–3937.

[10] D. F. Anderson and T. G. Kurtz. "Continuous time Markov chain models for chemical reaction networks". *Design and Analysis of Biomolecular Circuits* (2011).

[11] C. J. Oates and S. Mukherjee. "Network Inference and Biological Dynamics". *Ann. Appl. Stat.* **6**:3 (2012), pp. 1209–1235.

[12] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". *J. R. Stat. Soc. Series B Stat. Methodol.* **58**:1 (1996), pp. 267–288.

[13] R. Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective". *J. R. Stat. Soc. Series B Stat. Methodol.* **73**:3 (2011), pp. 273–282.

[14] W. Pan, Y. Yuan, and G.-B. Stan. "Reconstruction of Arbitrary Biochemical Reaction Networks: A Compressive Sensing Approach" (2012). arXiv: `1205.1720 [cs.SY]`.

[15] J. Wagener and H. Dette. "Bridge estimators and the adaptive Lasso under heteroscedasticity". *Mathematical Methods of Statistics* **21**:2 (2012), pp. 109–126.

[16] H. Zou. "The Adaptive Lasso and Its Oracle Properties". *J. Am. Stat. Assoc.* **101**:476 (2006), pp. 1418–1429.

[17] N. Meinshausen. "Relaxed Lasso". *Comput. Stat. Data Anal.* **52**:1 (2007), pp. 374–393.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". *Foundations and Trends in Machine Learning* (2011), pp. 1–123.

[19] N. Meinshausen and P. Bühlmann. "Stability selection". *J. R. Stat. Soc. Series B Stat. Methodol.* **72**:4 (2010), pp. 417–473.

[20] D. T. Gillespie. "Stochastic Simulation of Chemical Kinetics". *Annu. Rev. Phys. Chem.* **58**:1 (2007), pp. 35–55.

[21] J. G. Albeck, J. M. Burke, S. L. Spencer, D. A. Lauffenburger, and P. K. Sorger. "Modeling a snap-action, variable-delay switch controlling extrinsic cell death". *PLoS Biol.* **6**:12 (2008), pp. 2831–2852.

[22] Y. Chen, K. Shen, S.-O. Shan, and S. C. Kou. "Analyzing Single-Molecule Protein Transportation Experiments via Hierarchical Hidden Markov Models". *J. Am. Stat. Assoc.* ja (), pp. 1–49.

[23] P. Meyer, T. Cokelaer, D. Chandran *et al.* "Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach". *BMC Syst. Biol.* **8** (2014), p. 13.

[24] A. F. Villaverde, D. Henriques, K. Smallbone *et al.* "BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology". *BMC Syst. Biol.* **9** (2015), p. 8.

[25] J. J. Rice, Y. Tu, and G. Stolovitzky. “Reconstructing biological networks using conditional correlation analysis”. en. *Bioinformatics* **21**:6 (2005), pp. 765–773.

[26] S. Rogers and M. Girolami. “A Bayesian regression approach to the inference of regulatory networks from gene expression data”. en. *Bioinformatics* **21**:14 (2005), pp. 3131–3137.

[27] N. Friedman. “Inferring Cellular Networks Using Probabilistic Graphical Models”. *Science* **303**:5659 (2004), pp. 799–805.

[28] E. Segal, M. Shapira, A. Regev *et al.* “Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data”. *Nat. Genet.* **34**:2 (2003), pp. 166–176.

[29] S.-I. Lee, A. M. Dudley, D. Drubin *et al.* “Learning a Prior on Regulatory Potential from eQTL Data”. *PLoS Genet.* **5**:1 (2009). Ed. by A. M. Moses, e1000358.

[30] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. “Network medicine: a network-based approach to human disease”. en. *Nat. Rev. Genet.* **12**:1 (2011), pp. 56–68.

[31] M. Sunnåker, E. Zamora-Sillero, A. López Garcıa de Lomana *et al.* “Topological augmentation to infer hidden processes in biological systems”. *Bioinformatics* **30**:2 (2014), pp. 221–227.

[32] A. M. Klein, L. Mazutis, I. Akartuna *et al.* “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. *Cell* **161**:5 (2015), pp. 1187–1201.

[33] E. Z. Macosko, A. Basu, R. Satija *et al.* “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. *Cell* **161**:5 (2015), pp. 1202–1214.

[34] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. “Parameter Estimation for Differential Equations: A Generalized Smoothing Approach”. *J. R. Stat. Soc. Series B Stat. Methodol.* **69**:5 (2007), pp. 741–796.

[35] M. Dondelinger Fand Filippone, S. Rogers, and D. Husmeier. “ODE parameter inference using adaptive gradient matching with Gaussian processes”. *AISTATS* **31** (2013).

[36] P. Milner, C. S. Gillespie, and D. J. Wilkinson. “Moment closure approximations for stochastic kinetic models with rational rate laws”. en. *Math. Biosci.* **231**:2 (2011), pp. 99–104.

[37] T. Nakakuki, M. R. Birtwistle, Y. Saeki *et al.* “Ligand-specific c-Fos expression emerges from the spatiotemporal control of ErbB network dynamics”. *Cell* **141**:5 (2010), pp. 884–896.

[38] C. S. Gillespie. “Moment-closure approximations for mass-action models”. *Systems Biology, IET* **3**:1 (2009), pp. 52–58.

[39] J. M. Varah. “A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations”. *SIAM Journal on Scientific and Statistical Computing* **3**:1 (1982), pp. 28–46.

[40] A. A. Poyton, M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay. “Parameter estimation in continuous-time dynamic models using principal differential analysis”. *Comput. Chem. Eng.* **30**:4 (2006), pp. 698–708.

[41] B. Calderhead, M. Girolami, and N. D. Lawrence. “Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes”. *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Curran Associates, Inc., 2009, pp. 217–224.

[42] A. Ulianov, O. Müntener, and U. Schaltegger. “The ICPMS signal as a Poisson process: a review of basic concepts”. *Journal of Analytical Atomic Spectrometry* **30**:6 (2015), pp. 1297–1321.

[43] D. Grün, L. Kester, and A. van Oudenaarden. “Validation of noise models for single-cell transcriptomics”. *Nat. Methods* **11**:6 (2014), pp. 637–640.

3 Identification of mechanisms of fractional killing in TRAIL-induced apoptosis from mass cytometry time series snapshots

Anna Klimovskaia, Melissa Ko[1], Garry P Nolan, Manfred Claassen

Contribution by AK: conceived and implemented the computational analysis, interpreted resuls, wrote the manuscript.

---

[1] AK and MK contributed equally to this study

3.1. Introduction

Many pivotal cellular processes such as TRAIL-induced apoptosis are asynchronous, i.e. manifest themselves at rates that vary on a cell-by-cell basis. This variability constitutes a major challenge in learning the asynchrony-inducing mechanisms of these processes from experimental data. We propose RealMatch, an integrated single-cell biology and computational approach to address this challenge. We utilize RealMatch to derive insights into the mechanism for fractional killing in TRAIL-induced apoptosis.

Several tissues in the body are known to produce and secrete TNF-related apoptosis-inducing ligand (TRAIL), which generally induces apoptosis in tumor cells by binding to its respective receptors DR4/TRAIL-RI and DR5/TRAIL-RII. Prior work has investigated TRAIL for use as a anti-cancer therapeutic. However, rampant onset of drug tolerance in both in vitro and in vivo systems caused this therapy to fall out of favor. In-depth studies of in vitro systems such as HeLa cells have demonstrated the onset of non-genetic resistance upon exposure to TRAIL leading to a fraction killing effect. Even in TRAIL-sensitive cells, heterogeneity can be observed in the as evidenced by the asynchronous nature of apoptotic induction [1]. Differential responses to TRAIL has be attributed to multiple mechanisms, including but not limited to: genetic, epigenetic, or transcriptional alterations leading to changes in the protein levels of associated receptors, i.e. DR5, the death-inducing signaling complex (DISC) components, and apoptotic regulators like BCL-2 proteins or Inhibitors of apoptosis proteins (IAPs). Moreover, numerous signaling pathways such as NFkB, JNK, p38, and Erk have been implicated though evidence for their roles, pro-apoptotic or anti-apoptotic, have been contradictory. The precise roles of these many different pathways, especially across numerous individual cells with different levels of TRAIL sensitivity, still remain to be fully elucidated.

Studying an asynchronous, complex process involving many components is a difficult task. Capturing changes over time requires monitoring of the process, e.g. by means of a time course experiment. The asynchrony, i.e. the cell-to-cell variation in the process requires monitoring at a single-cell resolution. The complexity of the process given its many levels of regulation additionally requires a high dimensional single-cell approach that encompasses all important molecular process components, such as different signaling and apoptotic proteins in the case of TRAIL-induced apoptosis. A high dimensional time-lapse measurement would present the ideal experiment to study an asynchronous process. However, time-lapse measurements typically rely on reporter systems allowing tracking of very few components at the same time [2]. High-throughput single-cell snapshot measurements such as mass cytometry offer a suitable alternative to study synchronised multi-dimensional processes [3]. However, the destructive nature of the measurement allows assessment of the proteomic profile of each single cell only once, and in particular does not allow tracking of that individual cell over time. This situation leads to a challenge in reconstruction of the mechanisms of the underlying process from such data [4].

Several mechanistic models have been proposed to model the dynamics of TRAIL-induced apoptosis [5, 6]. However, these models unfortunately are not able to explain the fractional killing effect. While a mechanistic model of the fractional killing effect would be of interest in general, design of such a model for the case of a system with many unknown interactions and cross-talks between various pathways, is a tedious task. The state of the art approaches of model selection [7] would be too computationally expensive due to a vast majority of unknown interactions.

Conventional time series analysis such as Granger causality or Autoregression Models are not applicable to model single-cell behavior from snapshot measurements [8, 9]. These approaches are used to model the behavior of population averages over time and implicitly assume a homogeneous cell population at each time point. This assumption is violated for asynchronous process such as TRAIL-induced apoptosis. Asynchrony entails the appearance of heterogeneous cell populations at every time point with cells possibly covering the whole spectrum of states traversed in this process. Pseudotime ordering approaches address this issue to resolve the state sequence of dynamical processes from snapshot measurements of heterogeneous cell populations [10]. However, these approaches do not take into account time labels from time series experiments. Recently, an optimal transport model has been proposed to model developmental processes by matching cells between disjoint snapshots at different time points [11]. Any of these approaches reports a sequence of states, but does not yield the mechanisms governing the temporal transitions. This step has proven to be difficult and not yet solved satisfactorily [12]. In particular this is the case for the intricate task of inferring the mechanisms of fractional killing that enables a subset of cells to circumvent apoptosis upon external stimulation by TRAIL.

We propose RealMatch, an integrated experimental and computational approach to reconstruct asynchronous processes and demonstrate its ability to infer mechanisms underlying fractional killing in TRAIL-induced apoptosis. RealMatch operates on single-cell snapshot time course data and combines optimal transport modeling with approximative structure learning of dynamic systems to derive asynchrony-inducing mechanisms. We have carried out a mass cytometry time course study to monitor TRAIL-induced apoptosis and derive mechanistic insights into fractional killing. Further we validate this mechanism by evaluating the results through functional inhibitor experiments that perturb key pathways in our identified network.

3.2. Results

3.2.1. Analysis of TRAIL-induced Apoptosis by Mass Cytometry

We designed and carried out an experiment to inform about dynamics and emergence of cell subpopulations during TRAIL-induced apoptosis. Specifically, we collected HeLa cells at 30 minute intervals from baseline through six hours of TRAIL treatment analyzed by mass cytometry. We utilized an antibody panel including markers for signaling nodes encoding apoptotic state, apoptotic regulators, cell cycle stage, and signaling pathway activation (Table 1).

Mass cytometry profiling revealed several expected patterns in the time following addition of TRAIL, including induction of activated Caspase 3 (aCasp3) and cleavage of PARP (cPARP), as evidenced by an asinh ratio of nearly 3 in the median values at 6 hrs relative to baseline (**Figure 3.1A**). This increase in apoptotic state markers coincided with loss of full-length Bid caused by cleavage to produce the pro-apoptotic truncated Bid (tBid). On average, exposure to TRAIL resulted in a slight initial increase in signaling activation, i.e. an asinh ratio of 0.5 in p-NFkB, p-p38, and pErk median values, followed by a global downregulation of nearly all signaling pathways by three hours of TRAIL treatment, as reflected in an average asinh ratio of -1 to -3 in median expression values relative to baseline (**Figure 3.1A**).

However, recent work and our data has demonstrated that TRAIL-induced apoptosis is a highly asynchronous process where cells not only show difference in timing, but a fraction of cells does not irreversibly commit to programmed cell death. Observing the distribution of expression of several key markers confirmed this observation, as seen most strikingly in the bimodal distribution of Cisplatin staining, a reagent used in mass cytometry to stain dead cells (**Figure 3.1B**). Several other markers, including apop-

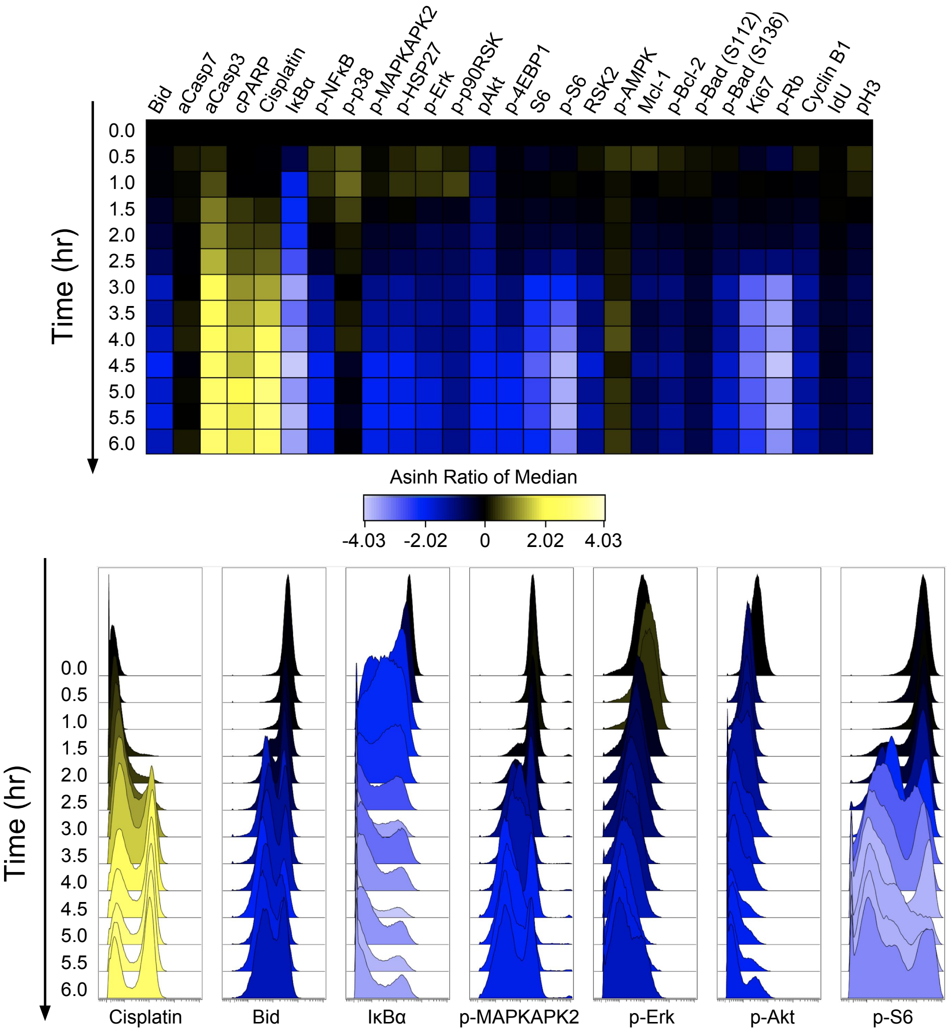


Figure 3.1.: Mass cytometric profiling of TRAIL-treated HeLa cells reveals bimodal response across time. (A) HeLa cells treated with TRAIL collected at 30 min. intervals across six hours were analyzed by mass cytometry. Heatmap shows asinh ratio of median expression intensities of each protein marker (columns) across all time points measured (rows), relative to the untreated baseline (first row). (B) Histograms displaying signal intensities of representative markers displaying emergence of bimodality with exposure to TRAIL.

totic regulators and signaling markers similarly revealed the emergence of different cell subpopulations from an apparently homogenous population at baseline. Of note, several markers such as IkBα, p-Akt, and p-S6 displayed a decreased median value at later time points of treatment, but the histograms revealed a subpopulation of cells that maintained nearly similar levels to that observed at baseline. This observation prompted our investigation to use this data to further infer the roles of these markers in driving persistence of a TRAIL-resistant subpopulation.

3.2.2. Inference of asynchrony inducing mechanisms from single-cell snapshot time series with RealMatch

Molecular mechanisms of the asynchrony underlying TRAIL-induced apoptosis are difficult to study and therefore are remain largely elusive. The destructive nature of mass cytometry results in losing the information about the "past" and the "future" of each individual cell in a snapshot. This gap of information constitutes a challenge to reconstruct the mechanisms of dynamic processes such as apoptosis, in particular those responsible for fractional killing. This challenge requires a general solution to trace back the initial cell subsets with future differential cell fate, such as those avoiding commitment to apoptosis and to infer the signaling network and in particular the driving interactions defining cell fate.

Snapshot measurements entail the appearance of heterogeneous cell populations at every time point with cells covering possibly the whole spectrum of states traversed in this process. Thus, understanding of this process requires a dedicated modeling approach to infer the molecular mechanisms of fractional killing. We propose RealMatch, a novel computational strategy to formulate hypothesis of potential molecular mechanisms of asynchrony in apoptosis from single cell snapshot time series data. This approach comprises three steps (**Figure 3.2**):

1. Definition of active signaling nodes in cell subsets with differential cell fate.
2. Inference of intracellular signaling network between the signaling nodes.
3. Identification of signaling mechanisms inducing differential cell fate.

For the first step, we reduce inferring the temporal dynamics of cell subsets to an optimal transport problem. This approach seeks to optimally "transport" a cell population state measurement from one time point into the measurement of the consecutive time point (**Figure 3.2A**). Optimality of transport is assessed with respect to a cost function, typically evaluating the distance of an individual cell from the previous time point to the state of another cell observed cell at the consecutive time point. This approach assumes a parsimony hypothesis, i.e. that cells favor changes with the least overall transportation cost. Therefore, by minimizing the cost function of total transportation cost between two samples, we can find an optimal "coupling" - a a set of transition matrices, each defining a probability of every cell at a specific time point to transition into the states represented by the cell of the consecutive time point. These transition matrices are subsequently used to reconstruct the maximally likely state trajectories from the start to the end of the process. These trajectories enable tracing back the initial cell subsets matching to those committed to apoptosis at a later time point as evaluated by elevated cell death (Cisplatin) signal. Combining this matching with the available status of a cell of being alive or dead at a given time, we can assign a label of time-of-death" to each trajectory and in particular to each cell at the onset of the time series (**Figure 3.2B**). Grouping of trajectories allows to evaluate the estimated temporal evolution of markers with respect to time-of-death and to identify potentially asynchrony-inducing markers characterized by differential behaviour across these groups (**Figure 3.2C**). To identify possibly non-trivial combinations of asynchrony-inducing signaling nodes we performed feature selection by regularized classification of initial cell states with respect to their inferred time-to-death (see Methods for details).

The second step consists in reconstructing the signaling network connecting the monitored signaling nodes. To this end we used Reactionet lasso, a structure learning algorithm for stochastic reaction networks monitored by a snapshot time series (see Methods for details). The outcome of Reactionet lasso procedure is a refined set of directed edges. Each inferred edge constitutes a potential regulation of an upstream protein to the downstream one. However, Reactionet lasso does not identify the functional role of the regulation in terms of an activation or inhibition. To define the functional role we associated a directed edge from protein A to protein B with an activating role if their time series are positively correlated, or an inhibiting role in case of negative correlation.

To achieve the third step, we analysed the network provided as an output of Reactionet lasso together with the "rainbow" plots from optimal transport for inhibition experiments (**Figure 3.2D**). This analysis allowed to reason what is the potential functional role of certain markers distinguishing surviving population and how manipulation on the abundance of these markers can affect cells' decision to die.

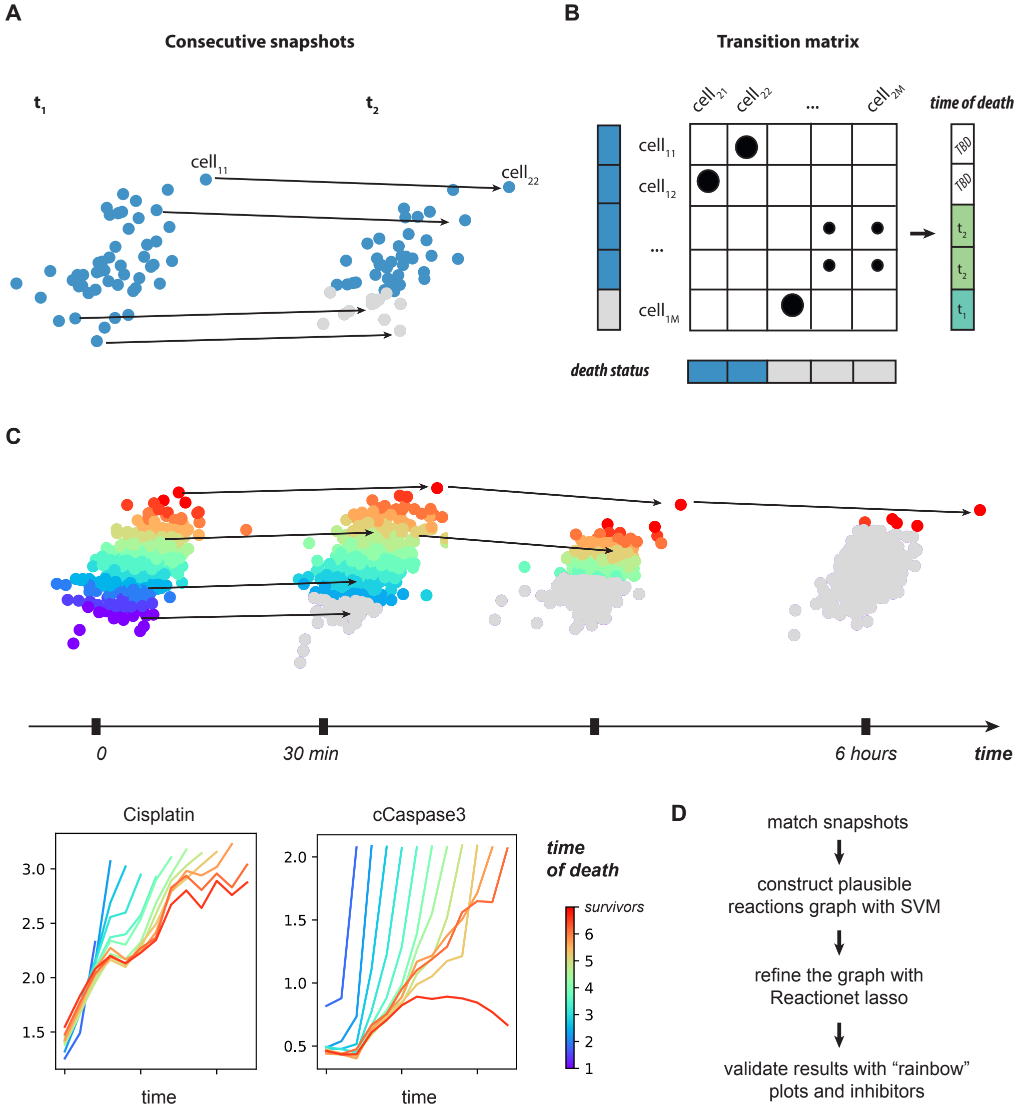

A
Consecutive snapshots
t_1
t_2
$cell_{11}$
$cell_{22}$
B
Transition matrix
$cell_{21}$
$cell_{22}$
...
$cell_{2M}$
time of death
$cell_{11}$
$cell_{12}$
...
$cell_{1M}$
TBD
TBD
t_2
t_2
t_1
death status
C
0
30 min
6 hours
time
Cisplatin
cCaspase3
time
time
time of death
survivors
6
5
4
3
2
1
3.0
2.5
2.0
1.5
2.0
1.5
1.0
0.5
D
match snapshots
construct plausible reactions graph with SVM
refine the graph with Reactionet lasso
validate results with "rainbow" plots and inhibitors

Figure 3.2.: Identification of mechanism of fractional killing from single-cell data with optimal transport and Reactionet lasso. (A) Optimal transport matching for two consequent snapshots. Blue and grey dots correspond to “alive” and “dead” cells accordingly. (B) Time of death reconstruction from the transition matrix and death status: if alive cell transitioned to the dead state we assign the corresponding time of death. (C) “Rainbow” plots reconstruction: cells are split into different subpopulations by the time of death and the median expression of the markers in each subpopulation is depicted. (D) Entire analysis pipeline: hypothesis formulation with optimal transport and reactions refinement with Reactionet lasso.

3.2.3. Identifying signaling nodes of survivor subpopulation after TRAIL Treatment

We applied RealMatch to identify key regulators and their role in fractional killing of TRAIL-induced apoptosis. First, we performed the optimal transport matching and consecutive marker analysis between different time-of-death groups. The evaluation of the average trajectory for each time-of-death group showed that for the majority of markers there is no significant difference between two consequent groups, but for some of the markers we see a significant monotonical increase in separation with the growth of time-of-death (**Figure 3.3**).

To quantitatively support these observations we trained a classifier to select a *survival profile*, i.e. significant subset of markers responsible for the classification of alive cells at the onset of TRAIL stimulation in two different future states: "committing apoptosis" and "survivors." The prediction accuracy of the corresponding classification data shows that there exist a subset of markers that allows to distinguish survival population (prediction accuracy 0.65) (**Figure 3.3**). Cell death markers like cisplatin do not achieve significant class separation at the onset of TRAIL stimulation, which agrees with the downstream role of these markers in apoptosis, i.e. occurring after cytochrome C release. As the expression of these markers increases over time, we see that the separation becomes significantly more apparent (**Figure 3.3**).

We focused on the markers which are detected to be significant at different time points before and after TRAIL exposure in order to understand the molecular profile of survivors (**Figure 3.3**). Markers that were identified as important in baseline, untreated cells included IkBα, p-BCL-2, pErk, and p-BADS136. IkBα was the most significant of these features (complement measure of significance). IkBα is a regulator of NFkB, a transcription factor which was previously reported to play an important role in TRAIL-induced apoptosis. However, the role of IkBα has not been clearly defined so far. Our results suggest that the levels of IkBα, rather than NFkB are important for fractional killing. Other significant markers include phosphorylated BCL-2 (p-BCL-2) and phosphorylated BAD (p-BADS136). This finding confirms previous studies reporting these players as indicators of inhibited cytochrome C release. In contrast to well-characterized anti-apoptotic markers, we identified active proliferation pathway kinase pErk as an important feature for cells that would not undergo TRAIL-induced apoptosis. Several markers became informative of survival status only after addition of TRAIL. For

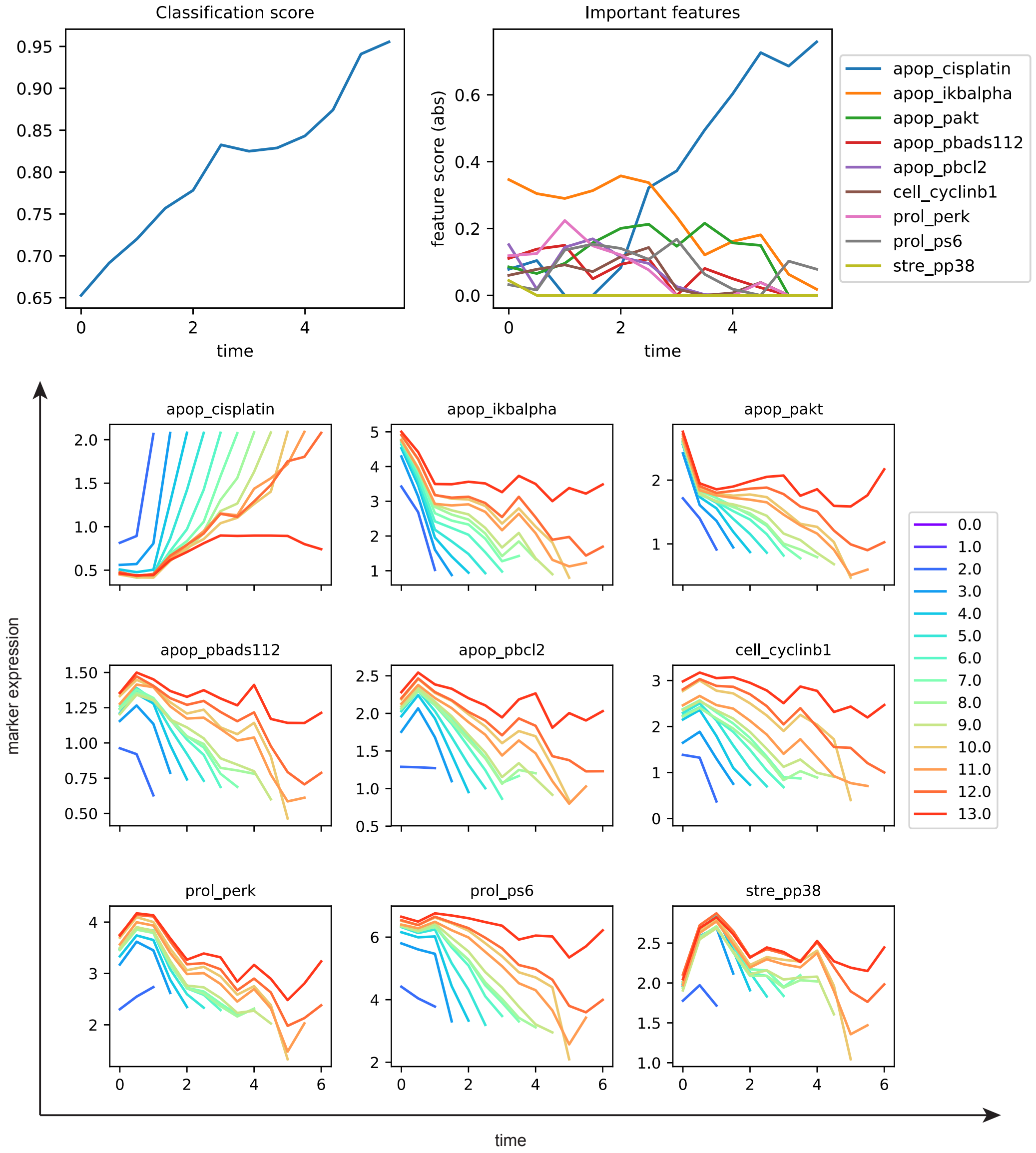


Figure 3.3.: Optimal transport matching identifies distinct trajectories of TRAIL-treated HeLa cells. The classification score shows steady increase in the classification accuracy between the apoptotic and survival populations. Feature score shows how significance of an individual feature changes over time. “Rainbow” plots show Asinh transformed median expression values of the most significant markers at the beginning of the process (t-0) in TRAIL-treated HeLa cells with different fates upon treatment with TRAIL.

example, translational regulator pS6, survival kinase pAkt, and the G2/M phase cell cycle marker CyclinB1 emerged as important apoptosis predictors closer to one hour of TRAIL treatment. By the final timepoints, pS6 was the only non-apoptotic marker (exclusive of Cisplatin) that remained informative. The role of many of these markers in TRAIL-induced apoptosis has not been previously reported. Though Akt participates in a well-known survival pathway, many of the other markers (e.g. S6, Erk, IkBα) have not been attributed a direct role in contributing to survival.

To validate that the *survival profile* plays a role in survival with TRAIL treatment, we tested targeted perturbations to assess the expected modulation of fractional killing. Given that IkBα or NFkB are regulated through degradation and localization mechanisms, robust and specific inhibition of these pathways generally requires genetic alterations not applicable in a clinical setting. Thus, we aimed to validate the rest of the edges that rely on kinase/phosphorylation dynamics. We used the p38 kinase inhibitor SB203580, the MEK inhibitor GSK1120212, and the PI3K inhibitor GDC0941. MEK is a kinase that phosphorylates Erk, leading to increased levels of pErk, while PI3K is an upstream kinase of pAkt. We tested these inhibitors in combination with TRAIL stimulation through clonogenic assays to assess their effects on long-term survival and proliferative recovery in HeLa cells (**Figure 3.5B**). Pre-treatment of HeLa cells with these inhibitors increased sensitivity to TRAIL exposure, i.e. with greater fractional killing demonstrated across all clonogenic assays. Interestingly, the MEKi demonstrated the greatest reduction (nearly 60% mean decrease in TRAIL survivors relative to DMSO) in TRAIL resistance of the three single-agent combinations tested.

3.2.4. Identification of potential mechanisms of TRAIL-induced apoptosis regulated by the identified key nodes

In order to explain the functional role and potential mechanism of action of the *survival profile* we performed Reactionet lasso analysis. We investigated potential reactions involving members of the *survival profile*, possibly explaining the efficacy of their inhibitors. There are several known regulators of cytochrome C release, specifically the BCL-2 family proteins (e.g. BID, MCL-1, BCL-2, BAD). While evidence suggests that various kinase pathways may play a role in regulating apoptosis, the relationship is not direct, suggesting that the markers we identified must interact with known apoptotic regulators.

We constructed a set of plausible reactions involving these markers and known regulators of cytochrome C release. We also added potential interactions between apoptotic, stress and proliferation pathways, which were not described before. Well-known interactions from literature we introduce as fixed reactions in the network in order to avoid structural unidentifiability and improve the performance. From the output (**Supplementary Figure B.1**) of Reactionet lasso we selected several potentially interesting reactions to investigate that were largely not explained in previous literature.

As expected, we identified crosstalk between measured BCL-2 family proteins such as connections between pBCL-2, pBADS112, and MCL-1. The contributions of IkBα to survival may be partially explained by the relationship between downstream pNFkB and the pAkt survival pathway or its connection to pro-survival protein MCL-1. Of interest, Reactionet lasso analysis implicated the p38 MAPK pathway as downstream of BID and leading to activation of pBCL-2 and pBADS112 via pHSP27. Moreover, our data suggests that many of these pro-survival BCL-2 family members, i.e. pBADS112, MCL-1, and pBADS136, feed into the pErk MAPK pathway.

As these reactions discovered by Reactionet lasso had not been reported before, we performed several additional inhibition experiment coupled with the mass cytometry measurements in order to validate and refine the proposed reactions. For experimental validation of the discovered reactions we followed a simple heuristic for causal inference: if we perturb a node in the network, we expect to see a significant change only in the downstream nodes. Therefore, based on our network inference we expect to see a significant change in dynamics (relative to DMSO control) of pErk under the inhibition of BCL-2 family and no significant change in the dynamics of BCL-2 under the inhibitions of pErk (**Figure 3.4**). In the case of p-p38, we expect perturbation BCL-2 proteins to perturb p38 signaling and the reverse to be true.

Using the most effective kinase inhibitors SB203580 and GSK1120212, we pre-treated HeLa cells prior to TRAIL exposure to measure the effect of targeting these markers through mass cytometry profiling. Moreover, as we had identified connections between these kinases and BCL-2 family members that might explain the means by which they regulate intrinsic apoptosis, we also collected mass cytometry data to characterize the effects of several BCL-2 specific inhibitors on TRAIL-induced apoptosis. From this data, we recreated the trajectories of cells towards different times of death or survival with combined treatment. This analysis reveals the change in the response of each individual marker per time-of-death group in order to answer if and when each marker was affected upon exposure to both a perturbation and TRAIL.

Relative to the DMSO control, MEKi and p38i reduced the characteristic spikes in both p-Erk and p-p38 that follow exposure to TRAIL (**Figure 3.4**). These kinase inhibitors increased apoptosis at earlier time points as seen with greater loss of BID at less than two hours of TRAIL treatment than what is seen in the DMSO control. Moreover, these kinase inhibitors also reduced levels of p-BADS136 and p-BCL-2, which recovered in late timepoints with p38i treatment, but dropped precipitously with MEKi treatment. Notably, this pattern suggests that the p38 and Erk MAPK pathways are able to influence levels of pro-survival BCL-2 family members or post-translational modifications. These observations can be partially explained by prior studies suggesting p90RSK kinase downstream of pErk feeding back into the BCL-2 proteins, namely pBADS112. Thus, these kinase inhibitors show potential to synergize with TRAIL and lead to reduced TRAIL resistance.

We similarly tested the effects of ABT-199, a selective BCL-2 inhibitor, and ABT-737, a BCL-XL, BCL-2 and BCL-w inhibitor, by mass cytometry compared to that of our kinase inhibitors (**Figure 3.4**). Inhibitors of pro-survival BCL-2 proteins have been developed as anti-cancer agents due to their ability to sensitize cancer cells to existing therapies or initiate cell death as a monotherapy. As the BCL-2 family work mainly via protein-protein interactions, these inhibitors compete for the binding sites in pro-survival BCL-2 proteins and displace pro-apoptotic BCL-2 proteins. This action allows pro-apoptotic BCL-2 family members to bind to and facilitate BAX/BAK complex formation, ultimately enabling apoptosis to occur. Both inhibitors led to notable reductions in pBADS136 and pBCL-2, as well as a characteristic drop in BID levels that explained the ability of these inhibitors to lead to reduced survival after TRAIL treatment. As expected, both inhibitors also had dramatic effects on levels of p-Erk and p-p38.

3.3. Conclusion

Deeper understanding into what drives susceptibility versus resistance to TRAIL could make great strides into developing TRAIL into an effective anti-cancer therapeutic. Prior technological limitations have led to ambiguous or even contradictory conclusions on the roles of various signaling pathways in regulating TRAIL-induced apoptosis. In part, this confusion resulted from methodologies that averaged measurements across a diverse population such as those undergoing TRAIL-induced apoptosis in an asynchronous manner and even a mixed population of cells sensitive or resistant to TRAIL. Recent advances in single-cell platforms such as mass cytometry enable us to characterize these populations in-depth. Using RealMatch, we aimed to recreate trajectories of cells responding differently to TRAIL and discover the underlying network regulating these heterogeneous responses. The final network derived from known reactions and those identified from the application of RealMatch to our mass cytometry data is depicted in **Figure 3.5A**. For the annotation of the "activation/inhibition" shown in this network, we used logic rules described above and the "rainbow" plots from a previous step (**Figure 3.4**).

This pathway analysis may explain the mechanisms of action of the inhibitors we investigated as possible synergistic co-treatments with TRAIL. We tested two kinase inhibitors directed against MEK and p38 as well as two BCL-2 family inhibitors, ABT-199 and ABT-737, in combination with TRAIL. MEKi is supposed to reduce levels of p-Erk though inhibition of its upstream kinase MEK, but we also see that it affects levels of p-p38, p-BCL-2, pBAD136 and pBADS112. p38i inhibits p38 activity and leads to a modest decrease in p-p38 levels, but also clearly affects pErk, pBCL-2, and p-BADS136 levels. ABT-737 reduces levels of BCL-2 family members pBCL-2, p-BADS136, and p-BADS112 while ABT-199 affects p-BADS136 and p-BCL-2 more dramatically with less effect on p-BADS112.

The observed effects of treatment with the inhibitors are consistent with and can inform the networks produced through Reactionet analysis. MEK inhibition leads to downstream effects on pBADS112 via a known reaction through p90RSK, but our inhibitor data as well as less stringent Reactionet analysis supports a connection between Erk and p38 signaling. This connection could explain how MEKi is able to also drive decreased levels of pBCL-2. Similarly, less stringent Reactionet analysis suggests the reverse connection between Erk and p38 is also true, which is supported by the dramatic reduction in pErk observed upon p38i treatment. Surprisingly, negligible effects were observed on pBADS112 in either treatment scenario. However, pBADS136 was significantly affected by both kinase inhibitors. This connection might have been missed by our Reactionet lasso method, especially if the connection is mediated by several unmeasured markers. Alternatively, our less stringent Reactionet results identified a connection between p-p38 and pAkt, which is the known upstream kinase of pBADS136. Thus, p38i could indirectly reduce pBADS136 via pAkt and the connection between p38 and pErk might explain MEKi's similar ability to downregulate pBADS136 through the same axis.

Importantly, this discovered network may provide explain the degree of efficacy we

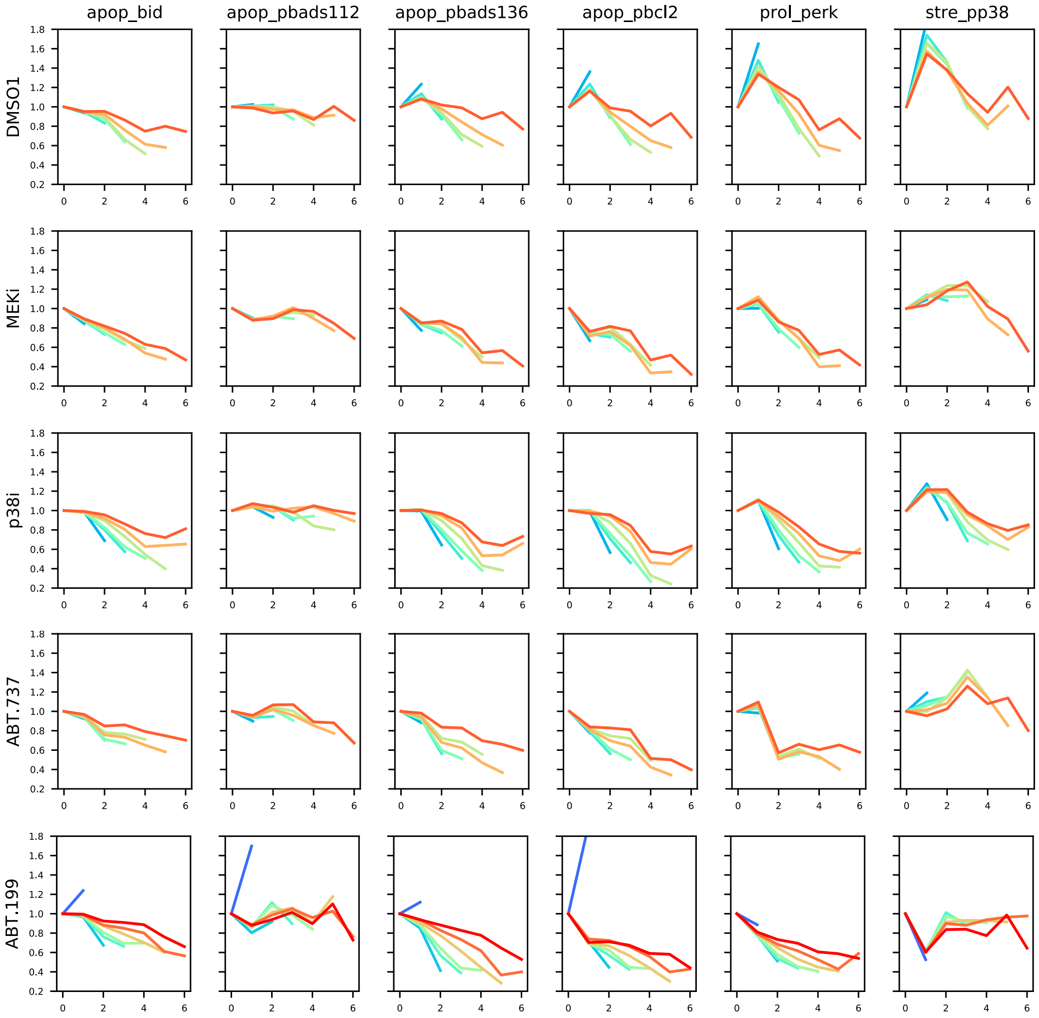


Figure 3.4.: “Rainbow” plots show Asinh transformed median expression values of some BCl-3 family markers, p-Erk and p-p38 in TRAIL-treated HeLa cells with various kinase inhibitors.

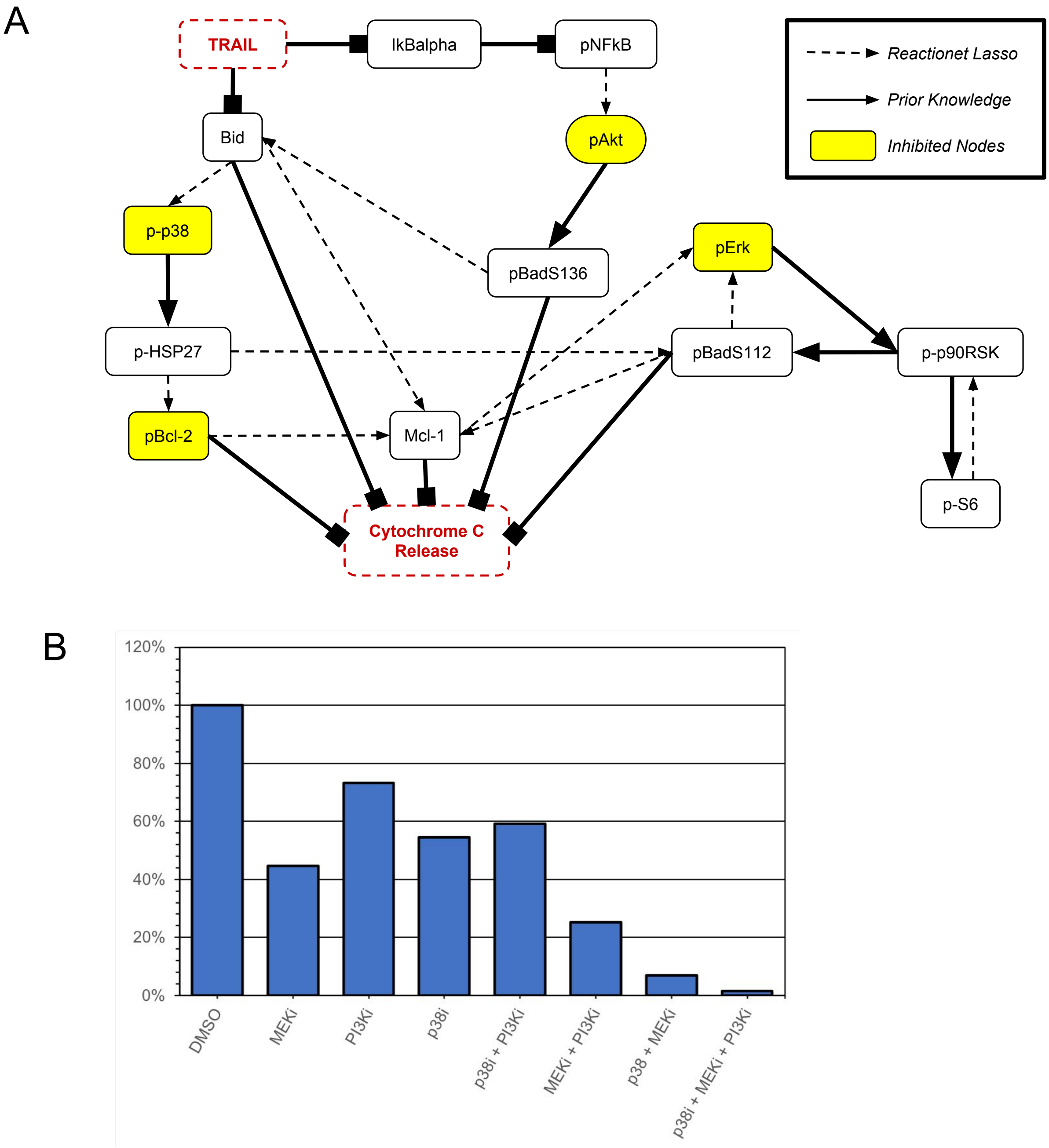


Figure 3.5.: Single-cell time course data implicates selectively maintained IkBalpha activation and downstream MAPK/survival pathways in TRAIL-resistant subpopulation. (A) Simplified Reactionet lasso graph documenting signaling network relationships revealed during TRAIL response. Known relationships documented in literature and novel relationships discovered from mass cytometry time course data are shown in solid and dotted lines respectively. (B) Histogram summarizing mean percentage of surviving clones following treatment with TRAIL plus targeted kinase inhibitors relative to TRAIL plus DMSO vehicle control, as measured by clonogenic assay.

observed with synergistic TRAIL co-treatments. We saw limited efficacy with the various BCL-2 family targeted inhibitors compared to kinase inhibitors despite conventional wisdom suggesting that the regulatory role of BCL-2 proteins in apoptosis would make them of greater import in co-treatments. We observed that the strongest death-inducing kinase inhibitors MEKi and p38i, which both reduced BCL-2 protein levels significantly through several proposed reactions that connect these pathways to measured pro-survival markers. Though anti-apoptotic regulation may be classically attributed to survival pathways such as pAkt, the hierarchy we identified might justify the ordering of efficacy we found: MEKi, p38i, and PI3Ki (**Figure 3.5B**). pErk's role upstream of p38 and pAkt, while also downregulating a separate axis through pBADS112 led to the greatest effect across multiple BCL-2 family members, while the effects of each subsequent marker (p-p38 and pAkt) encompassed fewer of these BCL-2 proteins.

Of interest, the network we identified might not fully reveal the rationale behind the efficacy of each pair-wise combination of these kinase inhibitors. Though the network suggests that MEKi should be largely redundant with PI3Ki and p38i due to pErk sitting upstream of both targeted nodes, we observe additional reduction in TRAIL resistance in both dual-kinase targeting strategies. Possible mechanisms for this phenomena include feedback loops wherein targeting of pErk may lead to a compensatory response through Akt or p38, which is subsequently blocked by either inhibitor, leading to further death. Alternatively, as the connections we found between pAkt, pErk, and p-p38 were very subtle and were labeled low confidence edges, it is possible that there is only a partial effect on either pAkt and p38 when inhibiting pErk levels. Adding a targeted inhibitor against either may achieve the full inhibitory effect. This mystery could be unraveled through future validation, such as profiling TRAIL-induced apoptosis using single-cell RNA-sequencing data that has fewer hidden variables or profiling of the different inhibitor combinations with the same approach via mass cytometry.

Studying asynchronous processes from disjoint mass cytometry snapshots is a difficult problem. Providing rich source of information by means of high-dimensional measurements, mass cytometry is still missing the vital aspect of following cells in time-lapse. A naive approach reconstructs the temporal state sequence on the basis of averaging the single-cell data, implicitly making the assumption of a synchronous processes yielding homogeneous cell state distributions over time. This approach is confounded by asynchronous processes such TRAIL induced apoptosis. RealMatch addresses the problem of disjoint snapshots by finding the best match between consequent snapshots under a parsimony assumption.

There are several options for such a matching approach. Our choice of a conventional discrete optimal transport was motivated by several factors as: simplicity and interpretability, demonstrated success of Waddington-Optimal Transport for transcriptomics applications [11] and the fact that minimizing total transportation cost fits well with the biological assumption of the cells choosing an optimal path in terms of “the least effort”. Computational complexity, i.e. quadratic memory and run time requirements constitute a limitation of this choice. Therefore, we considered cell subsets of 10’000 for the reported computations. Subsampling could in general introduce biases in the matching procedure. We ruled out such biases in our study by assessing stability of our results under various random seeds. Another limitation of such a matching is the “curse of dimensionality”: computing the cost function with large number of dimensions could lead toward insignificant (random) matches. To account for the fact that not all of our cells were reliably matched, we suggest to use only the signal trend, i.e. the average expression per time-of-death group, instead of performing inferences on the individual trajectories.

We reconstruct cell state trajectories for cell subsets with differential apoptosis outcome and use these to reason about potential causal mechanisms yielding different time of death in TRAIL induced apoptosis. We cast this task into a feature selection problem for classification of time-of-death. We used linear SVC with l1 penalty in order to select a sparse set of predictors. Another choice could be logistic regression with l1 penalty. However, logistic regression is known to be more sensitive to the outliers and SVC is easier to extend to non-linear case (non-linear SVC) in order to inspect the best prediction score of selected features assuming that their interaction is not linear. No matter which feature selection method we use, selected features are not guaranteed to be causal and could be just confounded with the outcome. In this case any drug intervention on such a feature wouldn’t lead to a different process behaviour. In order to identify the potential causal role of selected features we applied Reactionet lasso. Reactionet lasso showed good performance in the reconstruction of dynamic systems for the cases of systems with sufficient amount of prior knowledge about network structure. Reactionet lasso is the only known to us method for automatic reaction reconstruction from purely observational single-cell time series snapshots. Other approaches to this task are conceivable, e.g. performing model selection among several different topologies of an ordinary differential equation system. This approach could potentially give a better performance (confidence in the edge), but is typically limited by its required computational resources for evaluation of a potentially large number of model variants. The amount of models to consider for the considered marker panel in this study easily amounts to billions of variants that in practice are not tractable.

Time-series single-cell snapshots have never been used before for the analysis of asynchronous process because of the aforementioned limitations. We present RealMatch to overcome these limitations by means of an automatic and systematic procedure to identify asynchrony inducing mechanisms. RealMatch demonstrated its ability to discover potential causes of asynchrony in TIA. Further, we expect this approach to enable other studies based on time-series mass cytometry snapshots to reconstruct the mechanisms of other asynchronous processes. For example, RealMatch could be applied to studies of cells exhibiting trajectories profiled through well-defined status markers such as cell cycle or differentiation state.

3.4. Methods

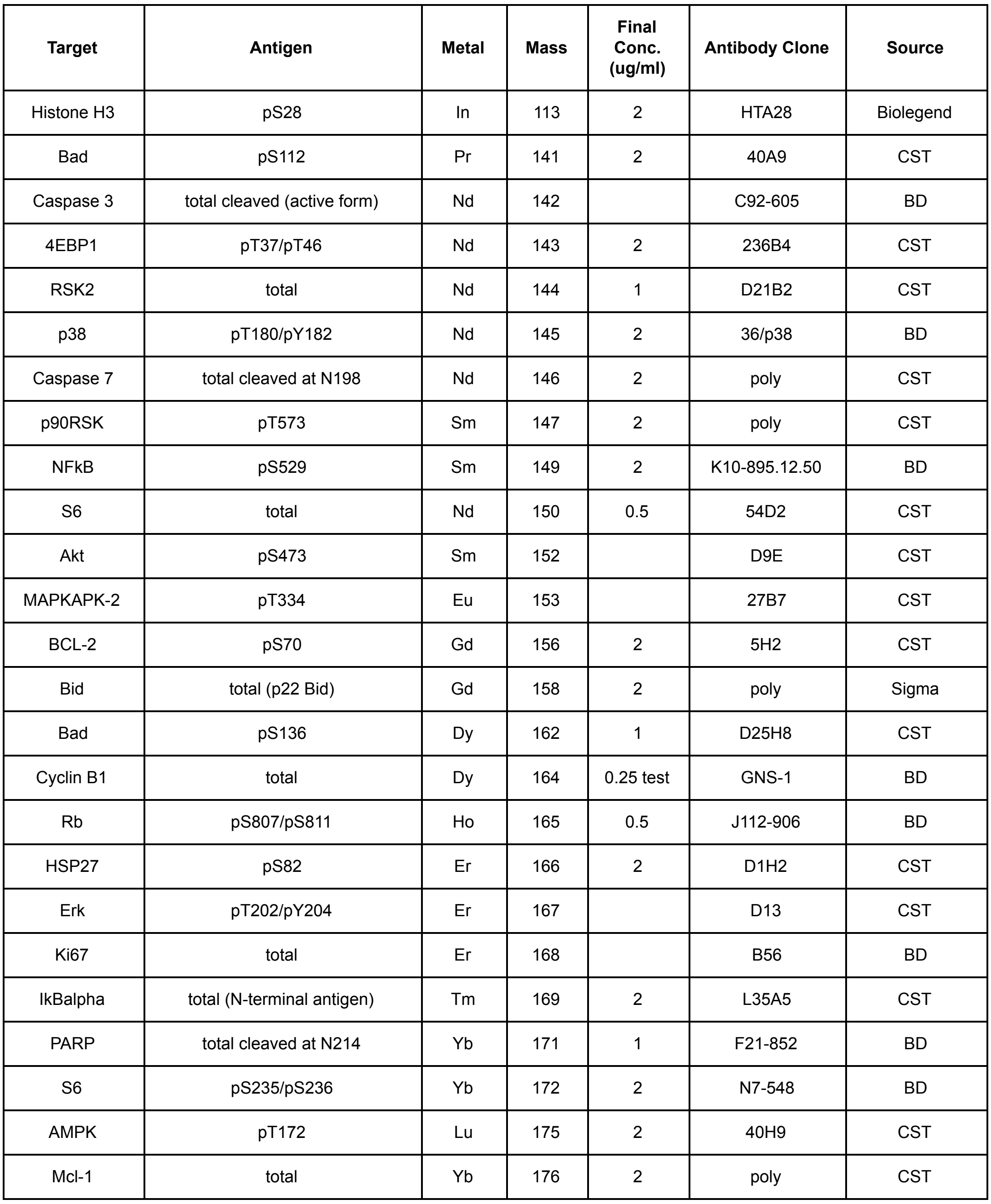
3.4.1. Reagents

Cell Culture and Inhibitors

The HeLa cell line was cultured in DMEM media containing 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin (Gibco). Cells were treated with 50 ng/ml SuperKiller TRAIL (Enzo Life Sciences). In inhibitor experiments, the following reagents and inhibitors were used: 0.1% DMSO (Sigma-Aldrich), 1 μM GSK1120212 (Selleck Chemicals), 20 μM SB203580 (Cell Signaling Technology), 2.5 μM GDC-0941 (Selleck Chemicals), 2 μM ABT-737 (ChemScene), and 2 μM ABT-199 (ChemScene).

Antibodies

The antibody panel including targeted epitopes, metal/mass conjugations, staining concentrations, clone, and manufacturers are described in detail below (**Figure 3.6**).



Target	Antigen	Metal	Mass	Final Conc. (ug/ml)	Antibody Clone	Source
Histone H3	pS28	In	113	2	HTA28	Biolegend
Bad	pS112	Pr	141	2	40A9	CST
Caspase 3	total cleaved (active form)	Nd	142		C92-605	BD
4EBP1	pT37/pT46	Nd	143	2	236B4	CST
RSK2	total	Nd	144	1	D21B2	CST
p38	pT180/pY182	Nd	145	2	36/p38	BD
Caspase 7	total cleaved at N198	Nd	146	2	poly	CST
p90RSK	pT573	Sm	147	2	poly	CST
NFkB	pS529	Sm	149	2	K10-895.12.50	BD
S6	total	Nd	150	0.5	54D2	CST
Akt	pS473	Sm	152		D9E	CST
MAPKAPK-2	pT334	Eu	153		27B7	CST
BCL-2	pS70	Gd	156	2	5H2	CST
Bid	total (p22 Bid)	Gd	158	2	poly	Sigma
Bad	pS136	Dy	162	1	D25H8	CST
Cyclin B1	total	Dy	164	0.25 test	GNS-1	BD
Rb	pS807/pS811	Ho	165	0.5	J112-906	BD
HSP27	pS82	Er	166	2	D1H2	CST
Erk	pT202/pY204	Er	167		D13	CST
Ki67	total	Er	168		B56	BD
IkBalpha	total (N-terminal antigen)	Tm	169	2	L35A5	CST
PARP	total cleaved at N214	Yb	171	1	F21-852	BD
S6	pS235/pS236	Yb	172	2	N7-548	BD
AMPK	pT172	Lu	175	2	40H9	CST
Mcl-1	total	Yb	176	2	poly	CST

Figure 3.6.: Antibodies used for mass cytometry experiments..

3.4.2. Clonogenic Survival Assays

To perform clonogenic assays, HeLa cells growing in log phase were seeded into 6-well plates at 2000 cells per well as counted by a manual hemacytometer about 18 hours prior to start of TRAIL treatment. Inhibitors or DMSO control was added the following day for 1 hour prior to TRAIL treatment, followed by the addition of TRAIL. Exposure to treatments continued for 24 hours before cells were washed twice with warm media and new media was added. Cell culture continued as described until colonies were visible to the naked eye and counted above 50 cells under the microscope, generally about 7-10 days later. At this point, media was removed from all 6-well plates. Colonies were stained by washing out remaining media using sterile PBS and fixing colonies using a cold methanol/acetone (60:40%) solution for 5-10 minutes. Colonies were visualized with a 1% crystal violet solution (dissolved in 25% methanol in water). Colonies on all plates were counted using a light box and counter pen (Fisher Scientific). All clonogenic assays are quantified as a percentage of surviving colonies relative to the TRAIL + DMSO control well.

3.4.3. Mass Cytometry

For mass cytometry experiments, HeLa cell line treatments were performed in 6-well plates using TRAIL and the inhibitors described under the Reagents section. Time course experiments were set up as a reverse time course with TRAIL treatment ranging from 0 to 6 hours sampled at every 30 minutes (TRAIL only experiment) or every hour (TRAIL plus inhibitor combinations). For TRAIL plus inhibitor treated cells, cells were treated with their respective inhibitor for 1 hr prior to addition of TRAIL. DMSO-treated cells were used as controls in all experiments with added inhibitors.

Prior to collecting time course samples, all cells were stained for cell cycle status with IdU and stained for viability with cisplatin according to previously published protocols. Cells from time course experiments were then lifted from plates using trypsin and fixed with 1.6% paraformaldehyde (Electron Microscopy Sciences) for 10 minutes at room temperature. Cells were pelleted and washed with cell staining medium (CSM) consisting of PBS with 0.5% BSA and 0.02% sodium azide. Cells were permeabilized with 4°p-NFC methanol for 10 minutes and stored at $-80°$.

Cells were barcoded using palladium barcoding as previously described to multiplex up to 20 different biological samples in a single experiment. After barcoding, cells were pelleted and washed three times CSM and stained with antibodies (Table 1) for one hour at room temperature. Cells were washed with CSM, then stained with 125 nm 191Ir/193Ir DNA intercalator (Fluidigm, South San Francisco, CA, USA) in PBS with 1.% paraformaldehyde at 4° overnight. Cells were washed once with CSM, washed three times with double-distilled water and filtered to remove aggregates. Prior to analysis, cells were resuspended with normalization beads and assayed on a CyTOF2 mass cytometer (Fluidigm, South San Francisco, CA, USA).

FCS files containing mass cytometry measurements were exported from CyTOF2 software. These files were concatenated, bead normalized, and de-barcoded using previously published software. Individual cells or singlets were gated on Cytobank (http://www.cytobank.org) using the event length and 191Ir/193Ir DNA parameters to remove debris and doublets.

3.4.4. Computational Approach

Data preprocessing

The data were preprocessed with a standard ashinh(x/5) transformation. Cells were classified into "alive" and "dead" with a strict threshold of 2.7 of the "cisplatin" marker.

Optimal transport

From each snapshot 10 000 cells were subsampled for the analysis. Python optimal transport library (`https://github.com/rflamary/POT`) was used to perform the matching. The transportation cost between each pair of cells in consecutive snapshots was computed as a squared euclidean distance with a cell cycle penalization:

$$c(x_1, x_2) = \begin{cases} ||x_1 - x_2||^2, \text{if } x_1 \text{agrees with the cell cycle } x_2 \\ +\infty, \text{otherwise} \end{cases} \tag{3.1}$$

By "agrees with the cell cycle" we mean that cells are not allowed transition back in the cell cycle (except from G1 to G0 phase) or skip cell phase (e.g. go from G0 directly to G2).

The output of the optimal transport between each pair of consequent snapshots is a transition probabilities matrix P containing probabilities of each cell of first snapshot transition into a cell state of the second snapshot. There for each cell of the first snapshot we can compute “expected state” as:

$$X^{\text{predicted}} = XP \tag{3.2}$$

For each cell we can estimate time-of-death as a first time when the predicted value of cisplatin for the “alive” cell in X became greater than the threshold used for the labeling last snapshot into “dead” and “alive”. Consequent matching of the snapshots allow to reconstruct the full trajectory from the beginning to the end of the process and assign expected time-of-death to each point of the zero snapshot. Cells with expected time-of-death 13 we call “survivors” meaning that they are alive by the moment of the last measured snapshot (for the total of 12 snapshots after TRAIL stimulation).

Feature selection

Linear Support Vector Classifier with L1 regularization was used to select the most informative features to classify cells into “survivors” and “committing apoptosis” at time point zero. The regularization parameter was chosen with a 5-fold cross validation. After the most significant features were selected, we rerun the SVC with “gaussian” kernel in order to access he best prediction score if the features could be non-linearly transformed.

Reactionet lasso

We ran Reactionet lasso with a standard set parameters proposed in the original paper: 100 bootstrap samples, splines gradient fit. We fixed 13 reactions as a prior stoichiometry based on the literature and allowed for 181 potential reactions. The set of potential reactions was chosen to connect important features identified in the step before between each other and to the known downstream regulators of apoptosis.

References

[1] S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. “Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis”. *Nature* **459**:7245 (2009), p. 428.

[2] S. Skylaki, O. Hilsenbeck, and T. Schroeder. “Challenges in long-term imaging and quantification of single-cell dynamics”. en. *Nat. Biotechnol.* **34**:11 (2016), pp. 1137–1144.

[3] B. Bodenmiller, E. R. Zunder, R. Finck *et al.* “Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators”. *Nat. Biotechnol.* **30**:9 (2012), pp. 858–867.

[4] K. Sachs, S. Itani, J. Fitzgerald *et al.* “Single timepoint models of dynamic systems”. en. *Interface Focus* **3**:4 (2013), p. 20130019.

[5] J. G. Albeck, J. M. Burke, S. L. Spencer, D. A. Lauffenburger, and P. K. Sorger. “Modeling a snap-action, variable-delay switch controlling extrinsic cell death”. *PLoS biology* **6**:12 (2008), e299.

[6] F. Bertaux, S. Stoma, D. Drasdo, and G. Batt. “Modeling dynamics of cell-to-cell variability in TRAIL-induced apoptosis explains fractional killing and predicts reversible resistance”. *PLoS computational biology* **10**:10 (2014), e1003893.

[7] M. Sunnåker, E. Zamora-Sillero, A. López García de Lomana *et al.* “Topological augmentation to infer hidden processes in biological systems”. *Bioinformatics* **30**:2 (2013), pp. 221–227.

[8] S. M. Hill, N. K. Nesser, K. Johnson-Camacho *et al.* “Context Specificity in Causal Signaling Networks Revealed by Phosphoprotein Profiling”. en. *Cell Syst* **4**:1 (2017), 73–83.e10.

[9] R. Opgen-Rhein and K. Strimmer. “Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process”. en. *BMC Bioinformatics* **8 Suppl 2** (2007), S3.

[10] K. R. Moon, J. S. Stanley, D. Burkhardt *et al.* “Manifold learning-based methods for analyzing single-cell RNA-sequencing data”. *Current Opinion in Systems Biology* **7** (2018), pp. 36–46.

[11] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary *et al.* “Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming”. *BioRxiv* (2017).

[12] A. Ocone, L. Haghverdi, N. S. Mueller, and F. J. Theis. “Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data”. *Bioinformatics* **31**:12 (2015), pp. i89–i96.

4 Causal learning of signaling pathways from single-cell time series snapshots

Anna Klimovskaia, Sara Magliacane, Stefan Ganscha, Fabian Radler, Manfred Claassen

Contribution by AK: conceived and implemented the computational analysis, designed and performed the computational experiments, wrote the manuscript.

4.1. Introduction

Signaling networks regulate ubiquitous cellular processes such as proliferation, differentiation and cell death [1]. The study of signaling networks typically centers around defining and taking advantage of causal relationships of its, typically protein components:

- Does protein X influence protein Y?
- If we increase/decrease/remove the abundance of protein X, how will it affect the final response of the signaling (e.g. arrest cell cycle, induce/inhibit apoptosis, etc.)?
- How to design an optimal treatment strategy for the system under study? (How to get the desired outcome, for example inducing apoptosis in cancer cells?)

Mechanistic modeling by means of dynamic systems is considered to be the most precise modeling approach in the attempt to answer these questions, for example ordinary and stochastic differential equations models are widely used to design new drug treatments [2]. Typical applications of mechanistic modeling assume the model structure to be known, which might not be true in practice. Discovery of new signaling interactions for a long time was precluded by the data: only few components were measured at a time, allowing to focus only on small parts of the pathway. Development of multi-parametric measurement technologies such as for example mass spectrometry or reverse-phase protein array time-course assays facilitated the development of algorithms to learn molecular interactions from data. The proposed approaches span from model selection for mechanistic models [3, 4] to recently statistical approaches extracting causal information from correlations and drug interventions [5, 6]. The major drawback of the aforementioned mechanistic approaches is that they are not scalable to large signaling networks. Even a more comprehensive enumeration of plausible network structures by topological augmentation is limited for the large scale systems with many unknown interactions [7]. Several regression based approaches were proposed recently to overcome the scalability limitation for learning mechanistic models from data [8, 9], but the structure learning capabilities could be largely precluded by overall identifiability of the models from the data [10]. The proposed statistical approaches in their turn typically don't suffer from computational limitations, but need a large variety of interventions to distinguish between correlation and causation.

Another important limitation of the aforementioned approaches for discovery of signaling interactions is that majority of them explicitly assume the signaling network to be deterministic. While this could be a reasonable approximation for certain type of systems, a number of recent studies with single-cell technologies, such as flow cytometry and time-lapse live-cell data, demonstrated that cell-to-cell variability plays a crucial role in systems on all levels from *intrinsic* variability (stochastic nature of molecular reactions) like in the examples of stochastic gene expression [11, 12] to *extrinsic* variability in protein concentrations during cell signaling, e.g. fractional killing effect in apoptosis [13]. Mathematical models explicitly addressing the heterogeneity demonstrate superior performance than deterministic approaches [14–17].

Development of new high-throughput single-cell measurement technologies, such as mass cytometry [18], allowed to measure up to 50 components in the system simultaneously and therefore unlocked a variety of opportunities for computational modeling of non-genetic origins of cell heterogeneity. However, addressing the cell-to-cell variability in structure learning task for mechanistic models still remains an extremely challenging problem [8].

Difficulty of structure learning for mechanistic models and the development of high-throughput single-cell technologies such as flow and mass cytometry facilitated application of statistical models in an attempt to generate hypothesis about causal protein interactions. For example, Bayesian Networks were proposed as a tool for causal hypothesis generation [19, 20]. These models essentially describe the joint distribution of the data as a tractable factorized distribution. Signaling relationships are represented as statistical dependencies implied by this factorization and can but in general do not coincide with the causal direction.

A more rigorous generalization of Bayesian Networks towards causal Bayesian Networks [21] and corresponding structure learning methods were recently proposed to infer causal structure from a single time point flow or mass cytometry snapshots of a signaling system in a steady-state [22–24]. However, even though causal Bayesian Networks indeed are representing truly causal dependencies, their applicability towards protein signaling is limited by two strong assumptions: the signaling system is at a steady-state and the signaling network is a directed acyclic graph (DAG) [25]. It is well-known that signaling systems contain a numerous amounts of feedback loops, which means the violation of the assumption of signaling network to be a DAG. In the latter study authors experimentally demonstrated that even in the case of perfectly designed experiment in terms of interventions, the structure learning capabilities of algorithms are indeed limited by the acyclicity assumption. Some recent theoretical work confirmed the significance of the aforementioned limitation by formulating the steady-state of a dynamic systems (mechanistic models) as a cyclic causal Structural Equation Models (SEM) for deterministic [26] and stochastic cases [27].

Dynamic Bayesian Networks would be one of the natural ways to overcome the problems imposed by single time point models [28]. However, structure learning methods for Dynamic Bayesian Networks [6, 29] are not directly applicable for flow or mass cytometry datasets due to their destructive nature: we don't posses the joint distribution over time snapshots. These models exploit temporal information of the process by formulating a time series model formulation. This information for cytometry datasets could only be acquired from multiple time snapshots and either by averaging or by matching of the snapshots (e.g. with optimal transport [30]). While averaging would lead to the loss of single-cell information and therefore not applicable to study cell-to-cell variability, the optimal transport could be an attractive alternative to make the Dynamic Bayesian Networks applicable for cytometry snapshot, but no study to our knowledge demonstrated it yet.

Despite aforementioned limitations (acyclicity) of recently proposed methods for causal structure learning from single time point cytometry snapshots, causal modeling is still very attractive and promising approach to modeling behaviour of complex signaling systems. First, because there are not many examples of success stories when causal modeling in biology was applied to learn gene regulatory networks [31–33]. Second, fast development of mathematical approaches for causal reasoning opens new opportunities to answer central systems biology questions we introduced in the beginning [34].

Here we introduce a Causal Time Series Snapshots Model (CTSSM), a special case of causal Structural Equations Models which allows to overcome the assumptions of the steady-state and acyclicity accounting for the time of the process. CTSSM also explicitly models *extrinsic* variability. The time component here is the backdoor between mechanistic and causal modeling: all the information is unrolled in time (no instantaneous interactions), resulting in the joint causal graph over time without cycles. This model formulation combines advantages of both mechanistic and statistical models: it is able to accommodate the most important structures in protein signaling such as feedback loops, variability in initial protein concentrations and temporal changes of node abundance, and at the same time takes advantage of the simplicity of the statistical models and

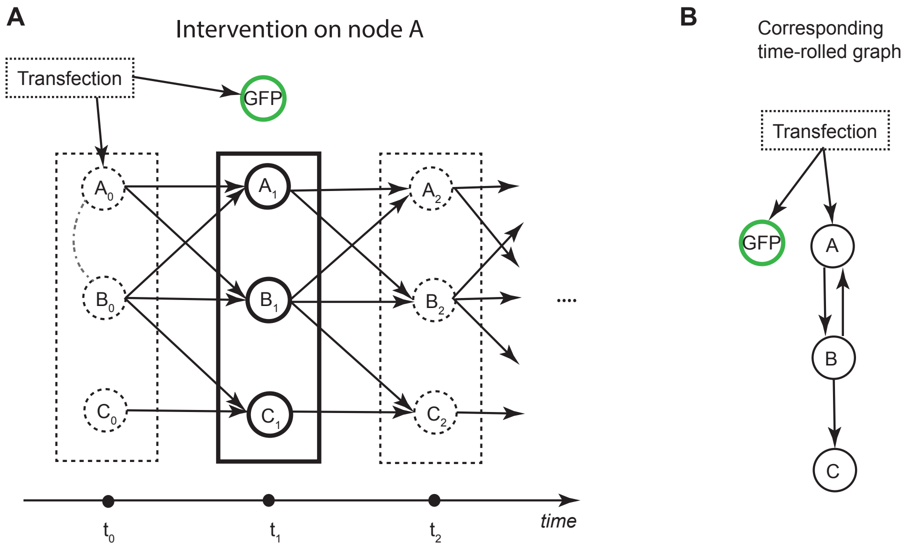

A
Intervention on node A
Transfection
GFP
A0
B0
C0
A1
B1
C1
A2
B2
C2
....
t0
t1
t2
time
B
Corresponding
time-rolled graph
Transfection
GFP
A
B
C

Figure 4.1.: General time series model for disjoint snapshots in the context of transfection experiments. **A** Time-unrolled model: individual box represents measured nodes per snapshot. We assume: (i) there are no instantaneous causal relations inside one snapshot, (ii) causal structure is preserved between pairs of consequent snapshots. Solid circles represent observed variables per snapshot. **B** Corresponding time-rolled causal model. Time-rolled graph is allowed to have cycles (negative and positive feedback loops).

directly transfers approaches developed for counterfactual reasoning in causal models.

We propose *MassCaRA* (Mass cytometry Causal Reaction Annotation algorithm) a simple method for practical discovery of signaling relationships in CSSTM from single-cell snapshots. We demonstrate performance of *MassCaRA* algorithm on several synthetic datasets, including one generated from a mechanistic model. By accounting for time, *MassCaRA* significantly outperformed aforementioned causal structure learning methods. We applied *MassCaRA* to a recently published dataset of EGFR signaling in HEK295T cells [35].

4.2. Results

4.2.1. Protein signaling as Causal Time Series Snapshots Model (CTSSM)

Consider a protein signaling network of n proteins measured at T distinct time points. We denote V a set of observed (measured) proteins. Lets assume a time series $\mathbf{X_t} = \left(X_t^i\right)_{i \in V}$ describes the corresponding protein signaling dynamics over time and satisfies the following assumptions:

- We denote $\mathbf{PA}^i$ a set of causal parents of protein i: intervention on any of the parents at any point in time will lead to a change of the distribution of protein i at later timepoints.
- We assume that causal structure is preserved between all the pairs of consequent snapshots (this assumption corresponds to the state of the art modeling of biochemical network as continuous time Markov chains) and there are no instantaneous signaling relations between variables in the same snapshot: $\forall i \in V$ there is a set $\mathbf{PA}^i \subseteq X^V$, s.t. $\forall t \in \{1, \ldots, T\}$

  $$X_t^i = f_i^t\left((\mathbf{PA}^i)_{t1}, X_{t-1}^i, N_t^i\right) \tag{4.1}$$

  with N_t^i (jointly) independent and for each i, N_t^i identically distributed in t and $F(X_0^1, \ldots, X_0^n)$ being a joint distribution of the nodes before the signaling was induced. N_t^i represents measurement noise per protein per snapshot.

One way to describe mechanistic models is to formulate them as discrete time systems [36]. Difference equations in discrete time systems are essentially Dynamic Structural Equations Models (DSEMs) [37], a special class of Structural Equations Models [21] unrolled over time. CTSSM is a DSEM for the case when only infrequent snapshots are observed and therefore in the case of infrequent mass cytometry time snapshots could be thought of as a crude approximation of the discrete time system.

The corresponding full time graph (**Figure 4.1A**) is obtained by drawing arrows from any node that appears in the right-hand side of (4.1) to X_t^i. As we don't allow any instantaneous effects the full time graph is always a Directed Acyclic Graph (DAG).

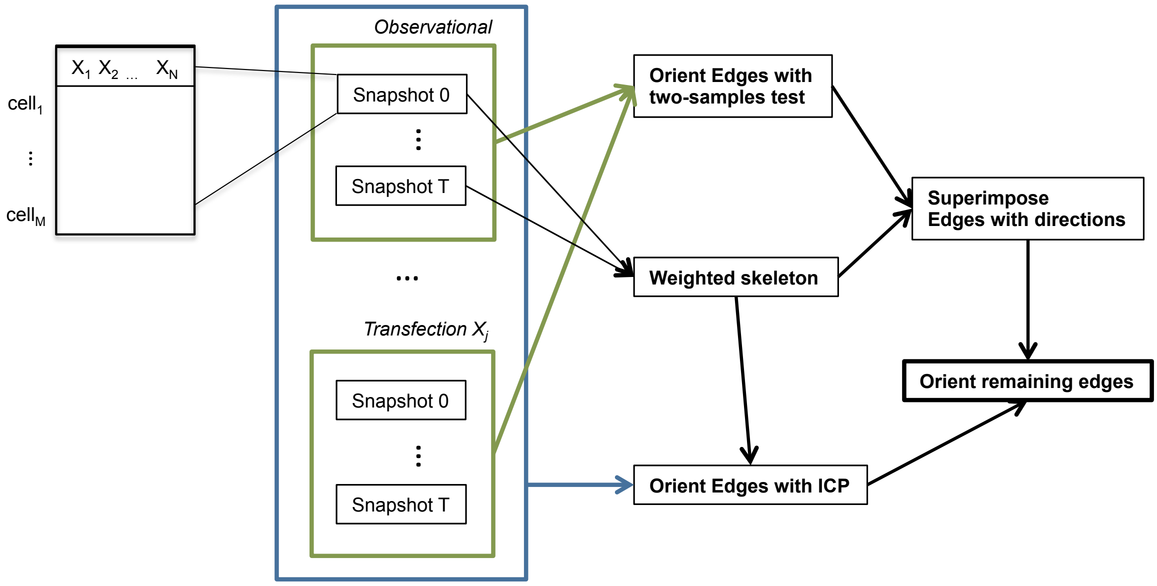

X_1 X_2 ... X_N
$cell_1$
⋮
$cell_M$
Observational
Snapshot 0
⋮
Snapshot T
...
Transfection X_j
Snapshot 0
⋮
Snapshot T
Orient Edges with two-samples test
Weighted skeleton
Superimpose Edges with directions
Orient remaining edges
Orient Edges with ICP

Figure 4.2.: *MassCaRA* pipeline. Green and blue boxes represent pooling of the data for different steps of the method.

Joint distribution over all time snapshots is always Markov[1] with respect to the full time graph (follows of it being a SEM). We additionally assume the joint distribution to be faithful[2] w.r.t. the full time graph. Therefore, CTSSM is on one hand just a special case of causal models introduced by Pearl [21] and we can use all the rich formalism developed for it: interventions, counterfactual analysis, causal inference; and on the other hand just a simple case of TiMINo model introduced by Peters et al.[38]. Unfortunately the corresponding methods for structure learning proposed by Peters et al. are not applicable for the cytometry snapshots since the joint distribution of the unrolled graph is not available.

By *intervention* in this study we denote any controlled manipulation on the value of a variable at any time point [21]. We must to point out that we clearly distinguish between "intervention" and "stimulation". In order to induce signaling, the system needs to be "stimulated", e.g. with a drug and only then we are talking about the signaling cascade. We don't consider stimulation as an intervention and assume the amount of stimulation to be preserved between different transfection experiments.

---

[1] The Markov condition states that each variable is probabilistically independent of its non-descendants given its parents in the graph.

[2] Faithfulness states that all the independence relations in the probability distribution over the variables are a consequence of the Markov condition.

The most common type of interventions considered in the literature are drug interventions, e.g. small molecular inhibitors, kinase inhibitors, etc. One major drawback of the interventions of this kind is that might be "unspecific": we might not know why the exact target and mechanism of these interventions. However, majority of causal structure learning methods assume require "surgical" interventions: the target and the value of the intervention is known precisely [21].

Though CTTSM and the structure learning approach proposed below are still valid for the case of drug intervention, in this study we focus on recently proposed interventions by transient transfection [35]. Mechanism of transfection is more transparent and easier to incorporate into CTSSM model than unspecific drug inhibitors. In case of transient transfection "intervention" is modeled as a change of the distribution of a target node at time point $t = 0$ (before the stimulation).

We denote by *experiment* a set of single-cell snapshots corresponding to one of the conditions: only stimulation and no intervention was performed we call ("observational"), and when some intervention on one of the nodes was performed ("interventional").

4.2.2. Mass Cytometry Causal Reactions Annotation (MassCaRA)

This section introduces the MassCaRA (**Figure 4.2**), a simple method for practical signaling relations discovery under the assumptions of CTSSM. As we mentioned before CTSSM satisfies TiMINo assumptions [38], therefore the joint distribution over all the snapshots is faithful with respect to the full time graph. In this case we could apply constraint-based causal discovery methods to recover the full time and corresponding time-rolled graphs. However, due to the destructive nature of mass cytometry we can only observe joint distributions inside one snapshot, but not essential for this approach joint distribution over all snapshots.

The cornerstone of constrained-based causal structure learning methods is d-separation[3] [21]. However, for CTSSM the essential set of nodes for d-separation is not available: to d-separate two nodes in one snapshot the d-separation set has to contain the nodes from previous time snapshots. Therefore, the structure recovered inside one snapshot using conventional constrained-based methods does necessarily contain all the ancestral casual relations and not only direct ones.

---

[3] two variables are d-separated if they are marginally dependent but become independent of each other once we condition on variables in d-separation set

Blom et al. [39] recently introduced the concept of "weak" d-separation to overcome the problem of measurement error. We propose to exploit the proposed "weak" d-separation as a heuristic approximation of the real d-separation as if we were conditioning on the right set of variables, e.g. we can only condition on B_t in the example from **Figure 4.1**, so A_t and C_t are not d-separated, but they could appear to be "weakly" d-separated (with a low confidence $\alpha < 10^{-6}$), which would allow us to remove an edge between them. Therefore, under "weakly" d-separation the recovered structure will be always more sparse. In the worst case this heuristic would select indirect causal relations instead of the direct one for the skeleton inference procedure, which is not dramatic for the skeleton, but could lead to misleading conclusions if it is used for edge orientation. Therefore, we propose to use it only for the skeleton learning part as a first step in MassCaRA. Also due to heuristic nature of the "weak" d-separation in this case, the learned skeleton could differ between different time points. We construct a joint skeleton between the snapshots by simple averaging of the edges in different snapshots. At the end of step 1 of MassCaRA we have a skeleton with a confidence score for each edge (details in Methods).

As a second step we acquire the edge direction with a simple straightforward procedure based on the definition of causality introduced above: for each node at each time point we test if its distribution changed significantly between control and intervention setting. In this case the edges will represent true, but only ancestral causal directions. We use the sign of the fold change in the directed edge in order to assign a type to the edge: "activation" (positive change) and "inhibition" (negative change). Unfortunately it is almost impossible to perform transfection experiment on every protein in the signaling network, therefore some edges will remain unoriented. To fill this gap we propose to use Invariant Causal Prediction (ICP) [40]. Successful applicability of ICP towards inferring causal relationships in genomics data with interventions by knock-outs was demonstrated before [33], therefore here we explore the performance of ICP for CTSSM.

As the last step of *MassCaRA* we superimpose the inferred directions and the skeleton favoring two-sample test directions above ICP in case of ambiguity. As we mentioned above, the final graph in the worst case contains all ancestral relations (direct and indirect). We suggest to use the weight of the edge from step 1 (skeleton) as a confidence in "closeness of relationships" (e.g. parents from grandparents and further ancestors): edges to direct parents are more likely to have a higher score.

4.2.3. Experimental results and benchmarks

Synthetic data

In this study we performed comprehensive comparison of the proposed method on several synthetic systems. We simulated 3 datasets of different levels of complexity from CSSTM, where the time-rolled graph is: a simple DAG ("EGFRsynDAG"), small network with two feedback loops ("EGFRsynCyc") and a large network corresponding to the available ground truth graph from Lun et al. ("EGFRsynLun") (**Supplementary Figure C.1**)). We also simulated a dataset from one of the state of the arts dynamic model (using ODE formulation and random initial conditions) for EGFR signaling [41] (**Supplementary Figure C.2**) in order to access performance of MassCaRA in a complicated case of mechanistic models (details in Methods).

We compared *MassCaRA* to two recently introduced methods for causal structure learning: *Invariant Causal Prediction* (ICP)[40] method and *backShift*[42]. Performance of ICP in biological setting [33] was demonstrated on the example of flow cytometry data with inhibitory interventions [22]. Performance of backShift [23] was demonstrated on mass cytometry data with various activators and inhibitors [43]. However, these settings are different from the dataset of Lun et al. in two major ways: they are based on a single time snapshot per experiment potentially corresponding to the steady-state of the signaling and majority of the interventions were performed with inhibitions, therefore constituting a different type of interventions ("surgical") opposed to the interventions with transient transfection ("fat-hand"). Both methods are build around a central concept of environment. Essentially environment is a different experimental condition. In case for example inhibition experiments each inhibition setting corresponds to the environment. On the other hand, different time snapshots could also serve as a different environments. Since methodologically it is not clear how we should handle the additional information coming from several time snapshots, we incorporated in three different ways:

- **Joint time.** We pull all the time snapshots from one experiment together to form one environment. In this case a sample from one environment is a sample from a joint distribution over time. Each environment corresponds to one experiment. Therefore, given M experiments, we get M environments.
- **Individual time.** We separate the problem into T sub-problems. In each problem we learn the structure from the distributions of "observational" and "interventional" experiments separately per time point and the total outcome average

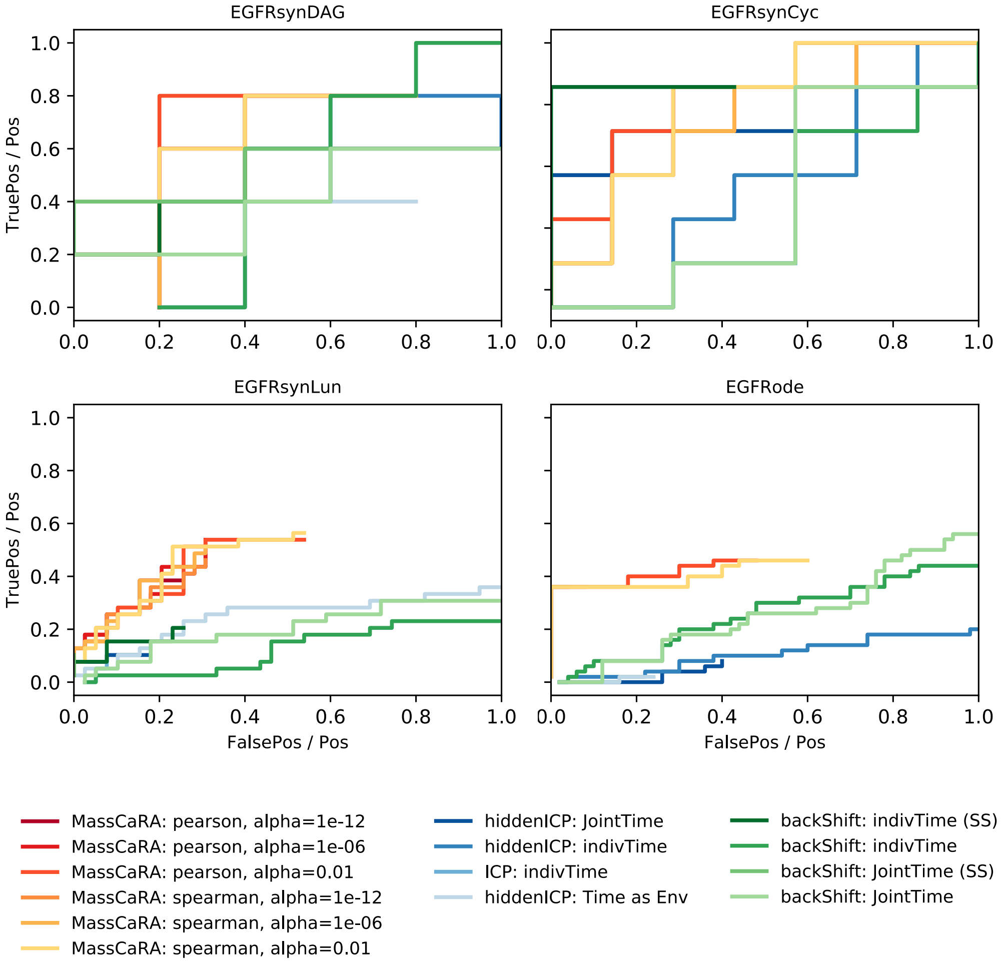


Figure 4.3.: Comparison of *MassCaRA* with different hyperparameteres against various settings of *ICP* and *backShift*. The comparison is done in terms of the ration of between true positives and false positives to the *total amount of positive edges*. We limited the x-axis to the value 1.0, because if the amount of false positives proposed by the method is greater than the total amount of positives, the results are not practical from biological point of view.

the joint edges. In this case we have T runs of a method with M environments in each.

- **Time as environment.** We treat each time snapshot in each experiment as a separate environment instance. Given T snapshots per experiment and M experiments we get $T * M$ environments.

The outcome of plain ICP was always an empty set for all setting, there we used the version of the methods with hidden variables (hiddenICP). As a metric of the performance we use a conventional true positive rate (fraction of correctly identified true positives to the total amount of positives) and a not commonly used fraction of positives to the total amount of positives. We don't use conventional Receiver Operator Characteristic curve since for perform *de novo* reconstruction of a sparse system and will mask the false positive signal due to a large amount of negatives. For practical biological applications such as consequent verification of the edges with experiments it is very important to achieve low amount of false positives since because of the cost of the experiments: only confident edges should go into validation [44].

Figure 4.3 demonstrates the results of the comparison. For the first two examples of test systems all the methods show good/comparable performance. This is largely due to the fact that the systems are too small and therefore don't contain too many indirect ancestral relations. Moreover, the models were generated from a Linear Structural Equations Model and therefore constitute an easy example for methods based on linear inference like ICP or backShift. Third example was generated from CSSTM with a system containing large amount of feedbacks. *MassCaRA* shows significantly better performance then ICP or backShift still containing relatively small amount of false positive reactions. For the dynamic systems case we didn't expect a very good performance of any of the methods since none of them specifically designed to handle structure learning of dynamic systems. However, we see that *MassCaRA* is consistently more correct than ICP or backShift.

Real data

Here we demonstrate application of *MassCaRA* to the EGFR signaling dataset from Lun et. al[35]. The data comprise several mass cytometry snapshots over signaling nodes measured at several different time points after the signaling stimulation and therefore constitute a perfect example of a CTSSM. 20 signaling proteins of HEK295T cells of

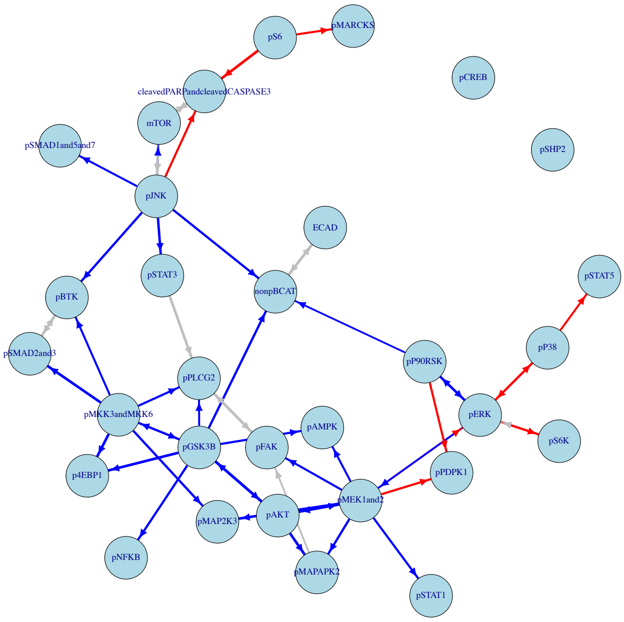

pS6
pMARCKS
pCREB
cleavedPARPandcleavedCASPASE3
mTOR
pSMAD1and5and7
pSHP2
pJNK
ECAD
pSTAT3
nonpBCAT
pSTAT5
pBTK
pSMAD2and3
pP90RSK
pP38
pPLCG2
pMKK3andMKK6
pERK
pAMPK
pS6K
pGSK3B
pFAK
p4EBP1
pPDPK1
pMEK1and2
pAKT
pMAP2K3
pNFKB
pMAPAPK2
pSTAT1

Figure 4.4.: Output of *MassCaRA* on Lun et al. dataset. Only edges with score above 0.6 were depicted for readability. Red edges correspond to positive activation and blue ones to negative feedback. Directions of red and blue edges was identified by transient transfection and gray edges correspond to the ICP directions.

EGFR signaling network were measured with a panel of 35 antibodies by mass cytometry during 1-h EGF stimulation time course. Measurements were performed at 0, 5 15, 30 and 60 minutes. Around 11 000 cells were measured per one experimental condition. A set of interventions was performed on the system by transient transfection: genes encoding protein of interest were cloned into vectors containing a cytomegalovirus (CMV) promoter and a GFP-tag sequence to transiently overexpress GFP-tagged proteins of interest. The tagged protein abundance was measured by mass cytometry using an anti-GFP antibody.

The original data were preprocessed with a standard $\text{arcsinh}(x/5)$ transformation. Cells in M phase of the cell cycle were removed from the study since cell cycle is known not to play a role in EGFR signaling yet cycling cells might introduce spurious correlations. For the skeleton inference procedure data were additionally preprocessed with the removal of outliers for each protein and the values below 5 counts were replaced with missing as it was suggested in the original study. All three replicates were used to obtain all the results in order to increase the robustness. For the skeleton inference it was implemented via median value in the covariance matrix. For the edge orientation the direction was acquired as a median direction between the replicates. Same applies for ICP. Parameters "spearman" and $\alpha = 10^{-12}$ were selected for the independence tests to achieve a sparse solution.

MassCaRA identified in total 84 edges. 70 of these edges were oriented based on transfection interventions. 34 edges correspond to positive interactions and 50 to negative feedback. In **Figure 4.4** we depicted 54 edges with a score above 0.6. 11 edges were reported previously in the SIGNOR database. We found 3 new interactions not reported in the SIGNOR database in particular interesting: positive feedback from pP38 to pERK, negative feedback from pP90RSK to pERK and negative feedback from pERK to pMEK1and2. Identification of these interactions wouldn't be possible without temporal information.

4.3. Methods

4.3.1. Learning sparse skeleton

First, the causal skeleton was obtained separately for each individual time snapshot from the observational experiment with a "skeleton" function from R package "PCALG" [45] and stability selection procedure [46]. Stability selection allows to increases robustness and obtain an estimation of a confidence score for an individual edge. Edges with a stability score less than 0.1 were removed. The final skeleton was constructed by obtaining an average score for each individual edge.

Skeleton inference procedure needs two parameters to be specified: a function for conditional independence tests and significance level $\alpha \in (0, 1)$ for the individual conditional independence tests. The standard choice for the conditional independence tests is based on "pearson" correlation. We additionally suggest to use "spearman" correlation for the case when the dependency between the variables is rather just monotonic, but not necessary linear. For the significance level a standard choice is $\alpha = 0.01$, but the lower levels of $\alpha = 10^{-12}, 10^{-6}$ would correspond to the "weak" d-separation.

4.3.2. Orient edges: "soft" interventions with transient transfection

Two-sample test

In order to obtain the direction of an edge from the skeleton, we performed non-parametric one-dimensional Kolmogorov-Smirnov two-sample test. Assume there is an edge in the skeleton between proteins A and B and transfection of protein A is available. Lets denote $B_0^{obs}, \dots, B_T^{obs}$ samples from one-dimensional distribution per individual snapshot $t_0, \dots, t_T$ for observational setting and $B_0^A, \dots, B_T^A$ for the corresponding transfection of A. For each $i \in \{t_1, \dots, t_T\}$ we test a null hypothesis of B_i^{obs} and B_i^A be drawn from the same continuous distribution and assign a weight to the edge $w^{A \to B} = \frac{1}{T} \sum_{i=1}^{T} 1_{H_0:\text{reject}}(\mathbb{E}B_i^A - \mathbb{E}B_i^{obs})$. We use a standard p-value 0.01 for the rejection of the hypothesis.

Invariant Causal Prediction

In order to obtain direction of the unseen interventions we use function "hiddenICP" from R package "InvariantCausalPrediction". ICP provides a set of potential causal ancestors for a target variable Y. Therefore, for each node in the graph we learn potential ancestors from observational and interventional setting (holding out an intervention on the node under study). Snapshots from different time points were pooled together into one environment for this study.

4.3.3. Simulating synthetic experiments

Simulations from CTSSM

For demonstration purposes that non-linearity is not the main issue the causal structure learning in this case all the datasets functional dependency was chosen to be linear. We first simulate distribution at $t = 0$ of every node i independently from $\mathcal{N}(a_i^0, s_i^0)$. At the next step we simulate "next time point" for every node i:

$$X_i^t = b_{ii}^t X_i^{t-1} + \sum_{j \in pa(i)} X_j^{t-1} b_{ij}^t, \tag{4.2}$$

where the $pa(i)$ - the set of causal parents of node i and the corresponding parameters for the distributions for $t = 0, \dots, T$ are sampled from: $a_i^t \sim \mathcal{U}[0, 2]$, $s_i^t \sim \mathcal{U}[1, 4]$, $b_{ij}^t \sim \mathcal{U}[1, 4]$.

In order to account to possible measurement noise, we model observed value as:

$$\hat{X}_i^t = X_i^t + \epsilon_i^t, \tag{4.3}$$

where $\epsilon_i^t \sim \mathcal{N}(0, 0.1)$ - measurement noise.

To model "fat-hand" interventions emulating transient transfection we simulate initial distribution from Gamma instead of Gaussian in order to account for heavy tails generated by transfection. We additionally introduce extra 'GFP-node' as a readout of the intervention.

$$\hat{X}_{GFP}^t = a_{GFP} X_{int}^0 + \epsilon_{GFP}^t, \tag{4.4}$$

for $t \in \{0, \dots, T\}$ and $\epsilon_{GFP}^t \sim \mathcal{N}(0, 0.1)$. We sample $a_{GFP} \sim \mathcal{U}[0.9, 1.2]$ to account for variability in the GFP read out.

For every setting we sample 6 time points, but use for the analysis only $t = 0, 1, 3, 6$ to model time intervals (0, 5, 15, 30 min) from Lun et al.

Simulations from mechanistic model

In this simulation we tried to reproduce experiment of Lun et al. with mechanistic modeling. We simulated a mechanistic model of "EGF and NGF signaling" [41] from an SBML model acquired from `http://www.ebi.ac.uk/biomodels-main/BIOMD0000000033`. The model was chosen to model phosphorylation mechanisms and to overlap in the most amount of nodes with the data measured by Lun et al. The model contains in total 32 nodes corresponding to both phosphorylated (active)/dephosphorylated(inactive) states of the protein. For the consequent analysis we assumed only phosphorylated (active) states to be observed. The reactions in the model are defined by Michaelis-Menten kinetics, and are reversible.

To reproduce mass cytometry single-cell experiment and to take in account destructiveness of the technology, we sampled all the snapshots separately. We assume the dynamics of the process to be deterministic, and we model single-cell variability with the variability in the initial conditions. The dynamics of individual cell was simulated with Ordinary Differential Equations with Matlab Simbio toolbox. Initial conditions were sampled randomly from log-normal distribution around original values provided in the SMBL model. All time points were sampled separately with reinitiate initial conditions. Time points were selected to correspond to 0, 5, 15, 30, 60 min intervals in the way that only last time point corresponds to the steady-state of the process. Interventions by transient transfection were modeled as a change of the initial distribution of the inactive form of the intervened protein from log-normal to a heavy tailed one. For the analysis the data were log-transformed.

4.4. Discussion

In this paper we introduced formulation of protein signaling as Causal Time Series Snapshots Model, a special case of causal dynamic Structural Equation Models accounting for infrequent time snapshots. The small modification by accounting for time compared to the classical single snapshot causal models, introduces a big advantage on the conceptual level: CTSSM allows to explicitly model feedback loops and cell-to-cell variability, the properties essential to apply the causal reasoning in a way it is traditionally performed by mechanistic models.

We introduced MassCaRA, a simple algorithm based on a combination of the constrained-based causal structure learning methods and recently proposed algorithm of Invariant Causal Prediction. MassCaRA consists of two parts: learning the skeleton and learning the causal directions. The skeleton learning part is based on a heuristic, however it shows a great performance on the synthetic data in order to refine arbitrary ancestral relations towards direct parents. This part of the procedure is generic and relies on purely observational data. The second part of edge orientation is theoretically sound and exploits model formulation as CTTSM. We demonstrated how "fat-hand" intervention could be used to obtain causal direction of the edges, but in principle any type of interventions could be applied here.

We demonstrated that MassCaRA significantly outperforms state of the art causal structure learning not only applied to models simulated from CTTSM, but to a mechanistic model generated from a dynamic system. The performance of MassCaRA on the mechanistic system suggests that the method could be well applicable for the real data scenario. When applied to Lun et al. dataset, *MassCaRA* revealed several potentially interesting feedback relations not reported in the SIGNOR database.

The demonstrated success of *MassCaRA* for structure learning for mechanistic systems, opens new horizons for the future research. First, an interesting direction for further investigation could be towards more explicitly accounting for time in the structure learning procedure. For example, optimal transport was recently proposed to overcome the problem of disjoint snapshots in single-cell transcriptomics data [30], however it's relation to causal structure learning is not established yet. Second important direction of the research is given a known CTSSM graph to infer the parameters of the model. Moment-based regression approaches could be for example used for this task [8]. And the last and the most interesting question for systems biology applications is to which extent in practice CTSSM could be used for causal reasoning, e.g. how to design drug interventions using CTSSM.

4.5. Acknowledgements

We would like to acknowledge Joris M. Mooij, Nicolai Meinshausen, James Wade and Vito Zanotelli for helpful discussions.

References

[1] H. Kitano. "Systems biology: a brief overview". *Science* **295**:5560 (2002), pp. 1662–1664.

[2] H. Kitano. "Computational systems biology". *Nature* **420**:6912 (2002), p. 206.

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

[4] T. Toni and M. P. Stumpf. "Simulation-based model selection for dynamical systems in systems and population biology". *Bioinformatics* **26**:1 (2009), pp. 104–110.

[5] R. O. Ness, K. Sachs, and O. Vitek. "From correlation to causality: Statistical approaches to learning regulatory relationships in large-scale biomolecular investigations". *Journal of proteome research* **15**:3 (2016), pp. 683–690.

[6] S. M. Hill, N. K. Nesser, K. Johnson-Camacho *et al.* "Context specificity in causal signaling networks revealed by phosphoprotein profiling". *Cell systems* **4**:1 (2017), pp. 73–83.

[7] M. Sunnåker, E. Zamora-Sillero, A. López García de Lomana *et al.* "Topological augmentation to infer hidden processes in biological systems". *Bioinformatics* **30**:2 (2013), pp. 221–227.

[8] A. Klimovskaia, S. Ganscha, and M. Claassen. "Sparse regression based structure learning of stochastic reaction networks from single cell snapshot time series". *PLoS computational biology* **12**:12 (2016), e1005234.

[9] Y. Pantazis and I. Tsamardinos. "A unified approach for sparse dynamical system inference from temporal measurements". *arXiv preprint arXiv:1710.00718* (2017).

[10] O.-T. Chis, J. R. Banga, and E. Balsa-Canto. "Structural identifiability of systems biology models: a critical comparison of methods". *PloS one* **6**:11 (2011), e27755.

[11] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. "Stochastic gene expression in a single cell". *Science* **297**:5584 (2002), pp. 1183–1186.

[12] B. Munsky, B. Trinh, and M. Khammash. "Listening to the noise: random fluctuations reveal gene network parameters". *Molecular systems biology* **5**:1 (2009), p. 318.

[13] S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. "Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis". *Nature* **459**:7245 (2009), p. 428.

[14] F. Bertaux, S. Stoma, D. Drasdo, and G. Batt. "Modeling dynamics of cell-to-cell variability in TRAIL-induced apoptosis explains fractional killing and predicts reversible resistance". *PLoS computational biology* **10**:10 (2014), e1003893.

[15] J. Ruess, A. Milias-Argeitis, and J. Lygeros. "Designing experiments to understand the variability in biochemical reaction networks". *Journal of the Royal Society Interface* **10**:88 (2013), p. 20130588.

[16] C. Zechner, J. Ruess, P. Krenn *et al.* "Moment-based inference predicts bimodality in transient gene expression". *Proceedings of the National Academy of Sciences* **109**:21 (2012), pp. 8340–8345.

[17] C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl. "Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings". *Nature methods* **11**:2 (2014), p. 197.

[18] S. C. Bendall, E. F. Simonds, P. Qiu *et al.* "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum". *Science* **332**:6030 (2011), pp. 687–696.

[19] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger. "Bayesian network approach to cell signaling pathway modeling". *Sci. STKE* **2002**:148 (2002), pe38–pe38.

[20] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. "Using Bayesian networks to analyze expression data". *Journal of computational biology* **7**:3-4 (2000), pp. 601–620.

[21] J. Pearl. *Causality.* Cambridge university press, 2009.

[22] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter single-cell data". *Science* **308**:5721 (2005), pp. 523–529.

[23] S. Triantafillou, V. Lagani, C. Heinze-Deml *et al.* "Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells". *Scientific reports* **7**:1 (2017), p. 12724.

[24] G. Papoutsoglou, G. Athineou, V. Lagani *et al.* "SCENERY: a web application for (causal) network reconstruction from cytometry data". *Nucleic acids research* **45**:W1 (2017), W270–W275.

[25] K. Sachs, S. Itani, J. Fitzgerald *et al.* "Single timepoint models of dynamic systems". *Interface focus* **3**:4 (2013), p. 20130019.

[26] J. M. Mooij, D. Janzing, and B. Schölkopf. "From ordinary differential equations to structural causal models: the deterministic case". *arXiv preprint arXiv:1304.7920* (2013).

[27] S. Bongers and J. M. Mooij. "From Random Differential Equations to Structural Causal Models: the stochastic case". *arXiv preprint arXiv:1803.08784* (2018).

[28] N. Friedman, K. Murphy, and S. Russell. "Learning the structure of dynamic probabilistic networks". *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc. 1998, pp. 139–147.

[29] C. J. Oates, J. Korkola, J. W. Gray, S. Mukherjee *et al.* "Joint estimation of multiple related biological networks". *The Annals of Applied Statistics* **8**:3 (2014), pp. 1892–1919.

[30] G. Schiebinger, J. Shu, M. Tabaka *et al.* "Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming." *BioRxiv* (2017), p. 191056.

[31] V. Lagani, S. Triantafillou, G. Ball, J. Tegner, and I. Tsamardinos. "Probabilistic computational causal discovery for systems biology". *Uncertainty in Biology.* Springer, 2016, pp. 33–73.

[32] M. H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. "Predicting causal effects in large-scale systems from observational data". *Nature Methods* **7**:4 (2010), p. 247.

[33] N. Meinshausen, A. Hauser, J. M. Mooij *et al.* "Methods for causal inference from gene perturbation experiments and validation". *Proceedings of the National Academy of Sciences* **113**:27 (2016), pp. 7361–7368.

[34] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect.* Basic Books, 2018.

[35] X.-K. Lun, V. R. Zanotelli, J. D. Wade *et al.* "Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry". *Nature biotechnology* **35**:2 (2017), p. 164.

[36] R. F. Stengel. *Optimal control and estimation.* Courier Corporation, 1994.

[37] T. Asparouhov, E. L. Hamaker, and B. Muthén. "Dynamic structural equation models". *Structural Equation Modeling: A Multidisciplinary Journal* **25**:3 (2018), pp. 359–388.

[38] J. Peters, D. Janzing, and B. Schölkopf. "Causal inference on time series using restricted structural equation models". *Advances in Neural Information Processing Systems.* 2013, pp. 154–162.

[39] T. Blom, A. Klimovskaia, S. Magliacane, and J. M. Mooij. "An Upper Bound for Random Measurement Error in Causal Discovery". *arXiv preprint arXiv:1810.07973* (2018).

[40] J. Peters, P. Bühlmann, and N. Meinshausen. "Causal inference by using invariant prediction: identification and confidence intervals". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**:5 (2016), pp. 947–1012.

[41] K. S. Brown, C. C. Hill, G. A. Calero *et al.* "The statistical mechanics of complex signaling networks: nerve growth factor signaling". *Physical biology* **1**:3 (2004), p. 184.

[42] D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. "BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions". *Advances in Neural Information Processing Systems.* 2015, pp. 1513–1521.

[43] B. Bodenmiller, E. R. Zunder, R. Finck *et al.* “Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators”. *Nature biotechnology* **30**:9 (2012), p. 858.

[44] S. Rogers and M. Girolami. “A Bayesian regression approach to the inference of regulatory networks from gene expression data”. *Bioinformatics* **21**:14 (2005), pp. 3131–3137.

[45] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, P. Bühlmann *et al.* “Causal inference using graphical models with the R package pcalg”. *Journal of Statistical Software* **47**:11 (2012), pp. 1–26.

[46] N. Meinshausen and P. Bühlmann. “Stability selection”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**:4 (2010), pp. 417–473.

Part III.

Concluding Remarks

5 Concluding Remarks

> *Everything should be made as simple as possible, but not simpler.*
>
> – Albert Einstein

Rapid development of single-cell proteomics measurement technologies, such as flow and mass cytometry, allowed to access the non-genetic origins of cell-to-cell variability simultaneously monitoring up to 30 markers at a time. In particular, with the new measurements, the heterogeneous response during protein signaling of a seemingly homogeneous population become apparent. Understanding the molecular mechanisms of cell-to-cell variability is crucial for design of treatment strategies for the diseases such as cancer [1].

There are two main sources of variability at a single-cell level: *extrinsic*, such as variability in the initial conditions before the signaling stimulation, and *intrinsic*, variability arising from the inherent stochastic nature of biochemical processes [2]. In this thesis, we discussed mathematical modeling approaches for both sources of variability. In particular, we focused on the question of reverse engineering these models (structure learning) from mass cytometry time snapshots for the cases when the majority of the signaling reactions is not known.

Chapter 2 addresses the question of learning topology of a Chemical Master Equation from data. Chemical Master Equation is a gold standard of mechanistic modeling of *intrinsic* variability by means of Stochastic Differential Equations. This chapter solely focuses on the problem of learning a topology of a reaction network in the presence of *intrinsic* variability. We introduced *Reactionet lasso*, a sparse regression based approach for *de novo* reconstruction of a topology of the Chemical Master Equation. *Reactionet lasso* is the first to our knowledge algorithm for automatic structure learning for Chemical Master Equation going beyond model selection.

We used the structure learning capabilities of *Reactionet lasso* to recover stochastic signaling interactions on synthetic data. This method is scalable and explicitly handles measurement noise potentially arising from the measurement. However, *Reactionet lasso* assumes all the components of a signaling network to be observed. Therefore, in practical application, when only part of the network could be measured due to the availability of antibodies, *Reactionet lasso* can not guarantee stable performance. Unfortunately there is no explicit way to account for unobserved variables in this approach, but the tests on synthetic systems showed that *Reactionet lasso* could achieve good performance in the presence of *sufficient* amount of prior knowledge. However, we can not quantitatively define what *sufficient* means in the context of an arbitrary signaling system; we suggest for practical applications keep a reasonable balance between the prior knowledge and plausible reactions.

Chapter 3 demonstrates an application of *Reactionet lasso* to uncovering the mechanisms of fractional killing during TRAIL-induced apoptosis. This chapter addresses both problems of *intrinsic* and *extrinsic* variability during signaling. The previously reported asynchrony of apoptosis after exposure to TRAIL can in principle arise because of two reasons: variation in protein concentrations before the onset of apoptosis and the stochastic nature of signaling reactions. Our collaborators performed extensive mass cytometry time snapshot measurements spanning the duration of TRAIL-induced apoptosis. The panel of measured markers was chosen to most informatively cover potential causes of apoptotic asynchrony in TRAIL-induced apoptosis. Analysis of individual snapshots demonstrated progressive dynamics of asynchrony: starting from a unimodal distribution before the stimulation of apoptosis (by the end of 6 hours) cells show distinct bimodal behavior.

Since *Reactionet lasso* only addresses *intrinsic* variability we combined it with snapshot matching approach by optimal transport. Cell matching between the snapshots is an indispensable part of addressing the *extrinsic* heterogeneity. Optimal transport allowed to select a set of markers, differences in initial concentration of which before the stimulus could be causing fractional killing effect. And as a next step *Reactionet lasso* was used to generate hypothesis in which signaling reactions these markers could be involved. Proposed mechanisms were validated with inhibitory experiments.

Despite the demonstrated success of application of *Reactionet lasso* towards understanding mechanisms of apoptotic signaling, it is a valid question if we were using "a hammer to crack a nut". The conclusions that we were drawing for the problem of fractional killing are rather of a *causal*, rather than mechanistic nature. In the end, apart of assuming the stochastic reaction networks to be a generating mechanism of intrinsic cell-to-cell variability, we never explicitly constructed a mechanistic model after. The main reason for that the the state of the art model of TRAIL-induced apoptosis [3] already contains around 60 nodes and doesn't include any interactions with stress or survival pathway which we demonstrated to be important. So if we would set ourselves with a task to construct a precise mechanistic model of the mechanistic model (mass action, Michaelis-Menten or Hill kinetics), we would be immediately lost in modeling latent nodes, which were not observed in our study. A legitimate question is if do we even need this detailed road map of the signaling in order to design drug treatment. If we zoom out a bit from modeling and just ask ourselves "why" do we need a mechanistic model, the answer would probably be "to design drug interventions". And if we have look in the simple hierarchy of modern approaches towards modeling systems [4], we will notice that the aforementioned question could be answered not only by means of mechanistic modeling, but with a perhaps simpler class of Structural Equation Models.

Structural Equation Models gained so far very little attention in Systems Biology, however their success to answer causal question in molecular biology was already recently demonstrated [5]. In **Chapter 4** we focus on a comprehensive modeling of *extrinsic* cell-to-cell variability with Structural Equations Models. We demonstrate that when the time explicitly incorporated into the Structural Equations Model (SEM), the feedback loops don't constitute a problem anymore. We propose *MassCaRA*, a simple algorithm to learn the structure of SEM from single-cell time snapshots and intervention experiments.

Obviously, there is "no free lunch" and the frequency of the time snapshots and the quality and amount of the interventions would affect the overall applicability of this approach towards signaling systems. Essentially, when performing snapshot measurements, we want to be sure that they cover the dynamics of the process. However, the latter statement is based on the scientific intuition and interesting question for the further research would be how to perform optimal experimental design in order to extract the most information about the system from experimental data. In particular, the questions would be: (i) how frequent should the snapshot be; (ii) would measuring the snapshots at equal time intervals help the structure learning procedure; (iii) which type interventions lead to a better identifiability of the model; (iv) how to infer not only the structure, but also the parameters of the model. Another interesting direction of the research would be given a Structural Equation Model, to design an optimal intervention strategy to achieve a desired outcome, e.g. complete induction of apoptosis avoiding fractional killing effect.

References

[1] S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. “Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis”. *Nature* **459**:7245 (2009), p. 428.

[2] A. Singh and M. Soltani. “Quantifying intrinsic and extrinsic variability in stochastic gene expression models”. *Plos one* **8**:12 (2013), e84301.

[3] J. G. Albeck, J. M. Burke, S. L. Spencer, D. A. Lauffenburger, and P. K. Sorger. “Modeling a snap-action, variable-delay switch controlling extrinsic cell death”. *PLoS biology* **6**:12 (2008), e299.

[4] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT press, 2017.

[5] N. Meinshausen, A. Hauser, J. M. Mooij *et al.* “Methods for causal inference from gene perturbation experiments and validation”. *Proceedings of the National Academy of Sciences* **113**:27 (2016), pp. 7361–7368.

6 Acknowledgements

First, I would like to say a big thank you to my supervisor **Prof. Manfred Claassen** for believing that I could do this big jump from mathematics to biology research. Thank you a lot for all you patience and support during this uneasy journey. I hope I learned to speak "biology language" if not at fluent, but at least at advanced level.

I would like to thank my committee **Prof. Uwe Sauer** and **Prof. Nicolai Meishnausen** for your genuine interest and valuable scientific input, that greatly helped me in this journey. And special thank you to **Nicolai** for introducing me to the world of Causality.

I would like to thank my collaborator **Dr. Melissa Ko** for being my guide in the world of molecular biology. It was very enjoyable to work with you. Thank you for your patience for answering my questions about apoptosis over and over again.

I would like to thank **Dr. Sara Magliacane** for all your support an Causality project. Your advices and corrections were invaluable to me and you patience to Causality is just contagious.

I would like to say a very special thank you to my friend **Dr. Valera Vishnevskiy** for all your support through all these years. Our discussion about life and science had a major influence on a formation of me as a researcher.

I would like to acknowledge all the past and present members of Claassen lab for a such an amazing working environment. The warm and friendly atmosphere in our group was helping to survive the moments when I already couldn't see the light at the end of the tunnel. Special thank you to **Will**, **Laura**, **Eirini**, **Ioana** for all your invaluable support.

Thank you my best ballet team **Gabi**, **Martha** and **Fumio**. You kept me sane and positive all these years, never allowing to give up. That's worth a lot.

I would like to thank my two best friends **Nastya** and **Valya** for all the patience with which you were supporting me all these years and all that optimism you were bringing to me.

I'm especially grateful to my parents for their love and encouragement.

Appendices

A Supplementary material Chapter 2

A.1. S1 Text

Moment expansion. Lets consider Stochastic Chemical Reaction Network described in *Methods*. Lets denote stoichiometric matrix of this system as $S = \{s\}_{ij}$. In mass action kinetics propensity function $a_l(\mathbf{x}; k_l)$ can be represented as $a_l(\mathbf{x}; k_l) = k_l * g_l(\mathbf{x})$, where

$$g_l(\mathbf{X}(t)) = \begin{cases} X_a(t)X_b(t), & \text{for bimolecular reactions } R_l : a + b \to \dots \\ X_a(t), & \text{for unimolecular reactions } R_l : a \to \dots \end{cases}$$

We can derive equations for moment expansion for mass action kinetics for moments of any order. However, for the reactionet lasso we use only moments up to second order, so we provide closed form equations only for these moments. Let denote $\mu_i(t) = \mathbb{E}[X_i(t)]$, $\sigma_i(t) = \mathrm{Var}[X_i(t)]$, $c_{ij}(t) = \mathrm{Cov}(X_i(t), X_j(t))$.

Then dynamics of means follows:

$$\frac{d\mu_i}{dt}(t) = \sum_l k_l * F_{il}(t), \tag{A.1}$$

where

$$F_{il}(t) = s_{il} * \mathbb{E}[g_l(\mathbf{X}(t))].$$

The dynamics of variances follows:

$$\frac{d\sigma_i}{dt}(t) = \sum_l k_l * F_{(N+i)l}(t), \tag{A.2}$$

where

$$F_{(N+i)l}(t) = s_{il}^2 * \mathbb{E}[g_l(\mathbf{X}(t))] + 2s_{il} * \mathrm{Cov}(X_i(t), g_l(\mathbf{X}(t))).$$

And the dynamics of covariances follows:

$$\frac{dc_{ij}}{dt}(t) = \sum_l k_l * F_{(N(i+1)+j)l}(t), \tag{A.3}$$

where

$$F_{(N(i+1)+j)l}(t) = s_{il}^2 * \mathbb{E}[g_l(\mathbf{X}(t))] + s_{il} * \mathrm{Cov}(X_j(t), g_l(\mathbf{X}(t))) + s_{jl} * \mathrm{Cov}(X_i(t), g_l(\mathbf{X}(t))).$$

A.2. S2 Text

Information criteria. The output of the reactionet lasso is a set of nested models. Selection of the optimal solution in terms of true/false positive tradeoff without ground truth still remains an open question. In the current paper we applied different strategies based on information criteria. To asses trade-off of goodness-of-fit and cardinality of the solution we used Akaike information criteria (AIC) or Bayesian information criteria (BIC).

A common selection strategy is to use the minimum of either of these functions. However, we found this strategy suboptimal for reactionet lasso. Evaluation of information criteria turns out to be monotonously improving for many problems, as it was the case for our problem instances. We therefore recommend to use 1st, 2nd etc. maximum of absolute change in the information criteria. Fig.2.6 demonstrates how absolute change of the BIC corresponds to desirable true/false positive trade-off (black dots labeled "1", "2", "3"). We applied described approach to the regression problem formulation from Eq.2.3 (see *Results*).

A.3. S3 Text

Inference of binomial noise correction for empirical moments. Lets X a random variable denoting theoretical abundance of a given specie s at a give time point t. Lets assume that as an outcome of an experiment we observe X_{obs}, which follows binomial distribution $Bi(X,p)$ with some fixed probability of success p. This allows us to explicitly formalize connections between moments of X and X_{obs}:

$$\mathbb{E}[X_{\text{obs}}] = \mathbb{E}\,\mathbb{E}[X_{\text{obs}}|X] = \mathbb{E}[Xp] = p\,\mathbb{E}[X] \tag{A.4}$$

$$\mathbb{E}[X_{\text{obs}}^2] = \mathbb{E}\,\mathbb{E}[X_{\text{obs}}^2|X] = \mathbb{E}[Xp(1-p) + X^2p^2] = p(1-p)\,\mathbb{E}[X] + p^2\,\mathbb{E}[X^2] \tag{A.5}$$

$$\mathbb{E}[X_{\text{obs}}^2] = \mathbb{E}\,\mathbb{E}[X_{\text{obs}}^2|X] = \mathbb{E}[Xp(1-p) + X^2p^2] = p(1-p)\,\mathbb{E}[X] + p^2\,\mathbb{E}[X^2] \tag{A.6}$$

$$\text{Var}[X_{\text{obs}}] = \mathbb{E}[X_{\text{obs}}^2] - [\mathbb{E}\,X_{\text{obs}}]^2 = \;= p(1-p)\,\mathbb{E}[X] + p^2\,\mathbb{E}[X^2] - p^2(\mathbb{E}\,X)^2 = p(1-p)\,\mathbb{E}\,X + p^2\text{Var}[X] \tag{A.7}$$

$$\mathbb{E}[X_{\text{obs, }1}X_{\text{obs, }2}] = \mathbb{E}\,\mathbb{E}[X_{\text{obs, }1}X_{\text{obs, }2}|X_1, X_2] = \mathbb{E}[pX_1pX_2] = p^2\,\mathbb{E}[X_1X_2] \tag{A.8}$$

$$\begin{aligned}\text{Cov}(X_{\text{obs, }1}, X_{\text{obs, }2}) &= \mathbb{E}[X_{\text{obs, }1}X_{\text{obs, }2}] - \mathbb{E}\,X_{\text{obs, }1}\,\mathbb{E}\,X_{\text{obs, }2} = \\ &= p^2\,\mathbb{E}[X_1X_2] - p\,\mathbb{E}[X_1]p\,\mathbb{E}[X_2] = p^2\text{Cov}(X_1, X_2)\end{aligned} \tag{A.9}$$

A.4. S4 Text

Biological replicates. As a default, we consider single time series experiments. Reactionet lasso analysis easily accommodates multiple replicates. Specifically, replicate specific response vectors b_k and design matrices A_k for each condition k are utilized to construct a problem instance by concatenation and apply reactionet lasso as described:

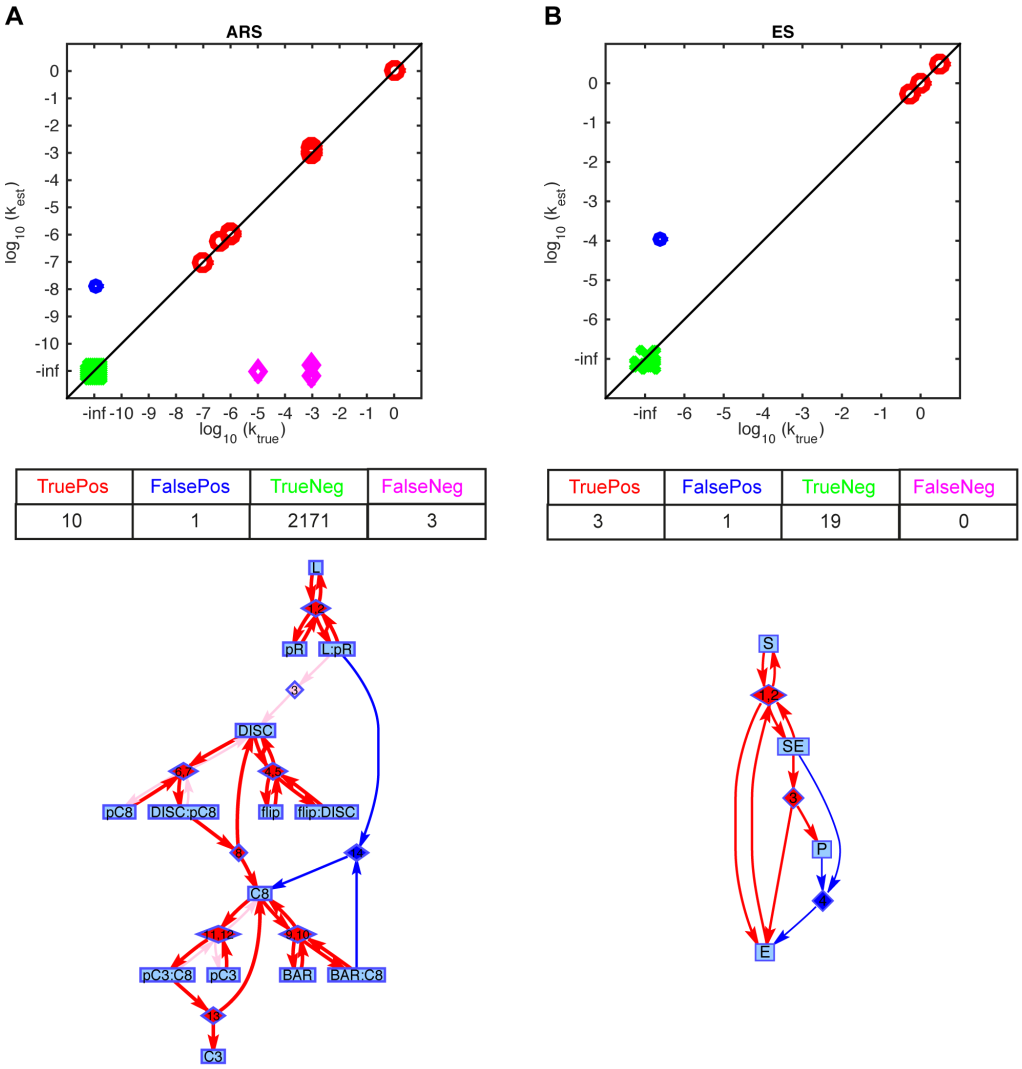
Matrix concatenation strategy:

1. Perform bootstrapping of stoichiometric moment function evaluations separately for each of the replicates k.
2. Construct joint regression problem by concatenation:
   response vector $b := [b_1 b_2 \dots b_n]$
   design matrix $A := [A_1 A_2 \dots A_n]$..
3. Run reactionet lasso on the joint regression problem defined by b and A.

A.5. S1 Dataset

For TRAIL induced apoptosis signaling cascade inference we used 33 equally distributed time points (in seconds) between 0 and 8 hours: [0, 900, 1800, 2700, 3600, 4500, 5400, 6300, 7200, 8100, 9000, 9900, 10800, 11700, 12600, 13500, 14400, 15300, 16200, 17100, 18000, 18900, 19800, 20700, 21600, 22500, 23400, 24300, 25200, 26100, 27000, 27900, 28800].

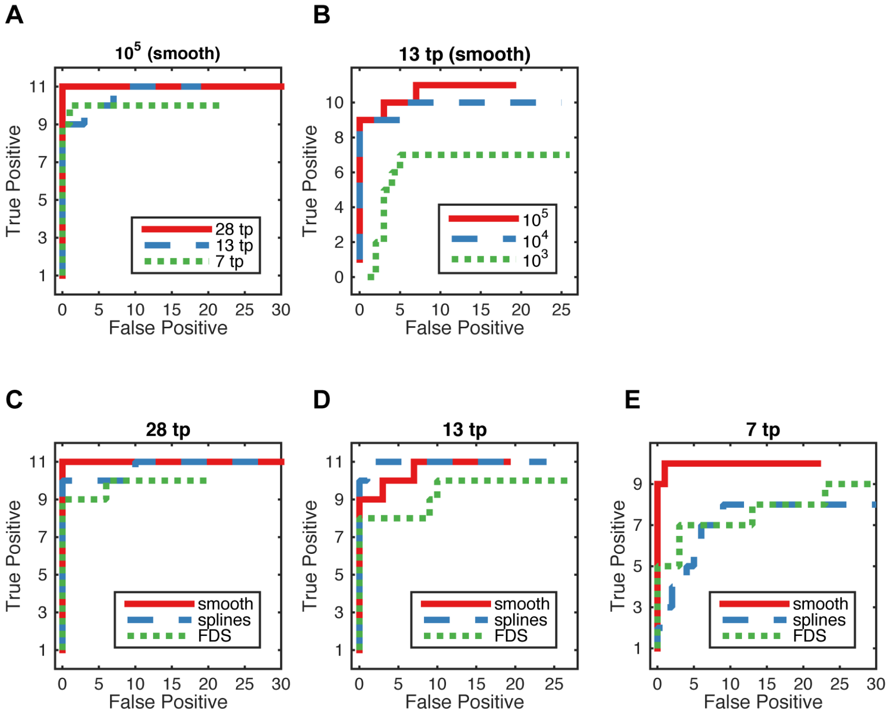
A.6. Supplementary Figures



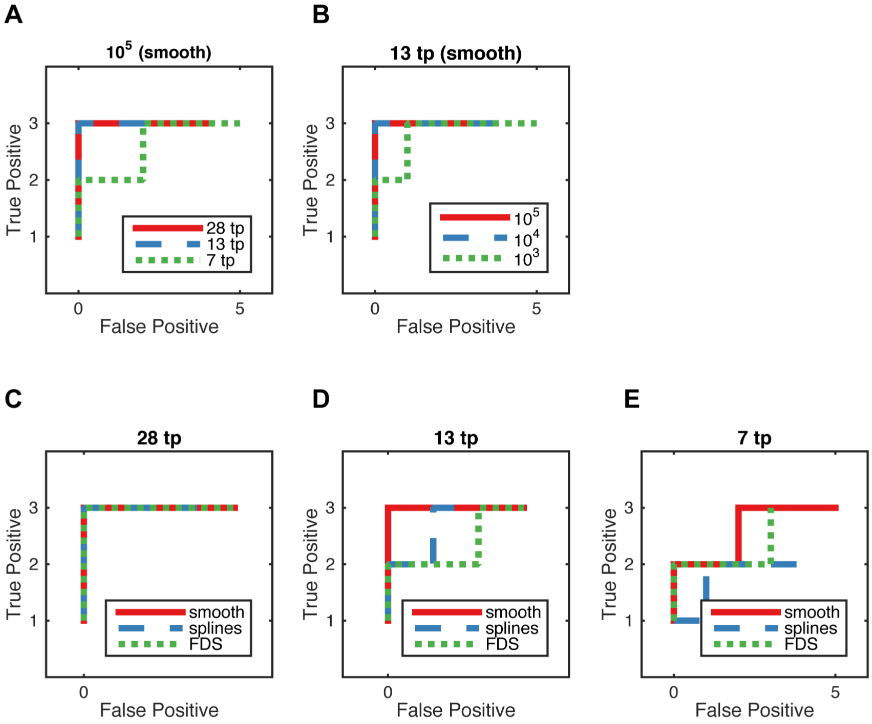
TruePos	FalsePos	TrueNeg	FalseNeg
10	1	2171	3

TruePos	FalsePos	TrueNeg	FalseNeg
3	1	19	0

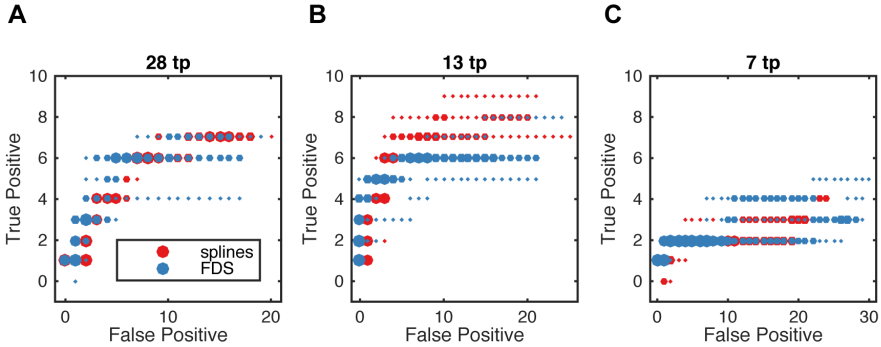
Supplementary Figure A.1.: Structure learning performance of the *reactionet lasso* for 10^5 single cell trajectories evaluated at 13 time points for **(A)** apoptotic receptor subunit (no measurement noise); **(B)** the enzymatic system. Empirical moment gradients estimated with cubic splines. Solution selected with Bayesian Information Criteria (BIC).



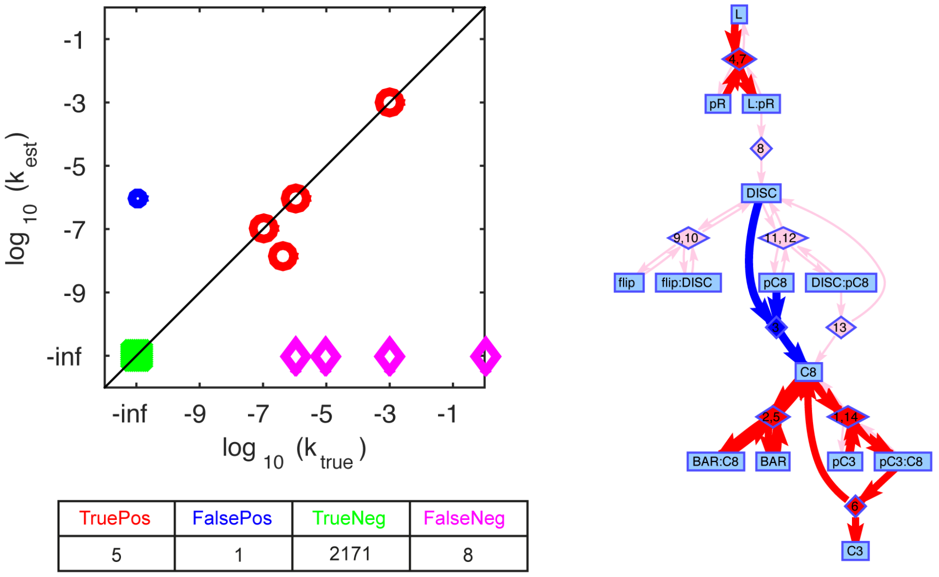
Supplementary Figure A.2.: Regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit (no measurement noise). **(A-B)** Empirical moment gradients estimated with “smooth” procedure: **(A)** 10^5 single cell trajectories evaluated at different amount of time points (tp) as indicated in the legend. **(B)** Different number of single cell trajectories: 10^3, 10^4, 10^5 evaluated at thirteen time points. **(C-E)** Results for different empirical moment gradient estimates: smooth(red), splines (blue), FDS (green) for different amount of time points: 28 (C), 13 (D), 7 (E).



Supplementary Figure A.3.: Regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to enzymatic system (no measurement noise). **(A-B)** Empirical moment gradients estimated with "smooth" procedure: **(A)** 10^5 single cell trajectories evaluated at different amount of time points (tp) as indicated in the legend. **(B)** Different number of single cell trajectories: 10^3, 10^4, 10^5 evaluated at thirteen time points. **(C-E)** Results for different empirical moment gradient estimates: smooth(red), splines (blue), FDS (green) for different amount of time points: 28 (C), 13 (D), 7 (E).

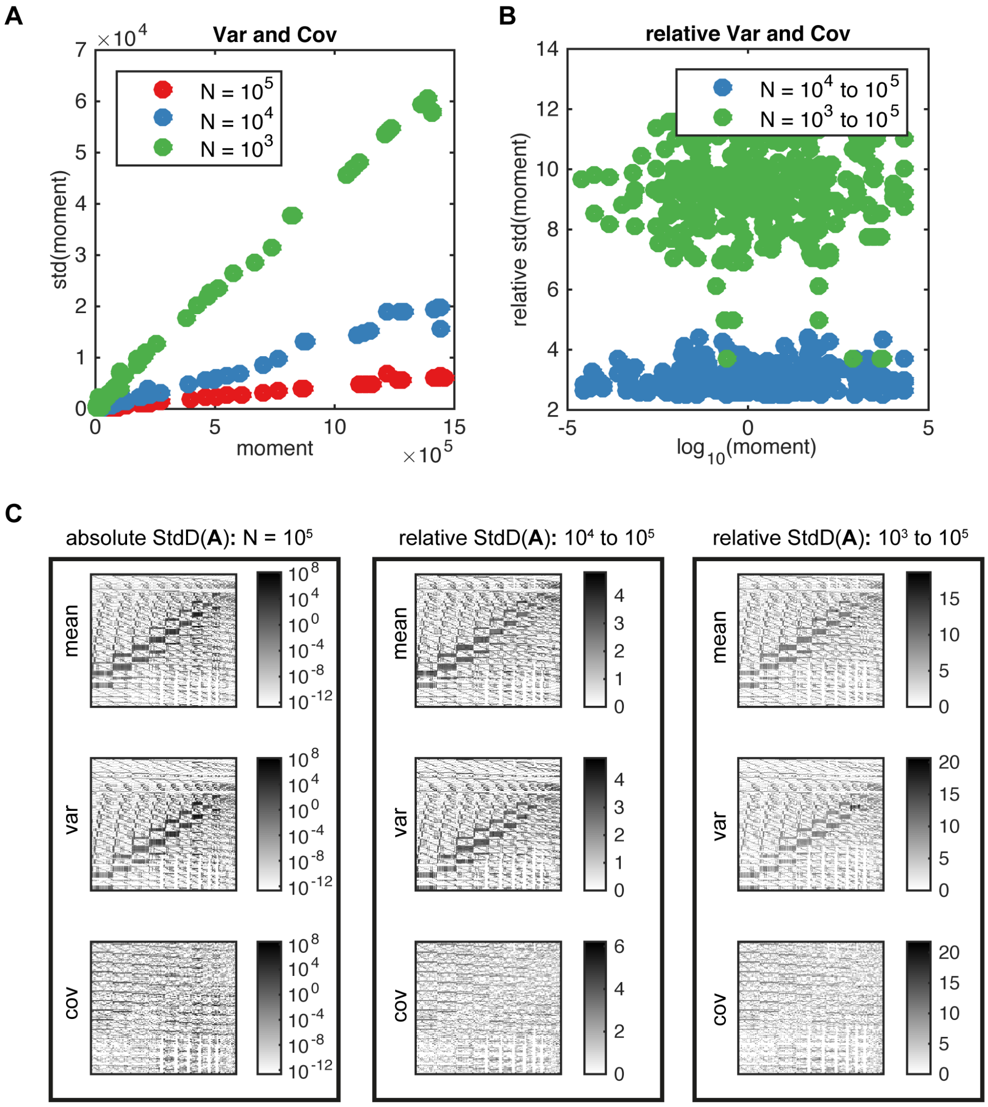


Supplementary Figure A.4.: Overlay of five regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit (p = 0.05) with 13 time points. Results for different empirical moment gradient estimates: splines (red), FDS (blue) for different amount of time points: 28 (A), 13 (B), 7 (C).

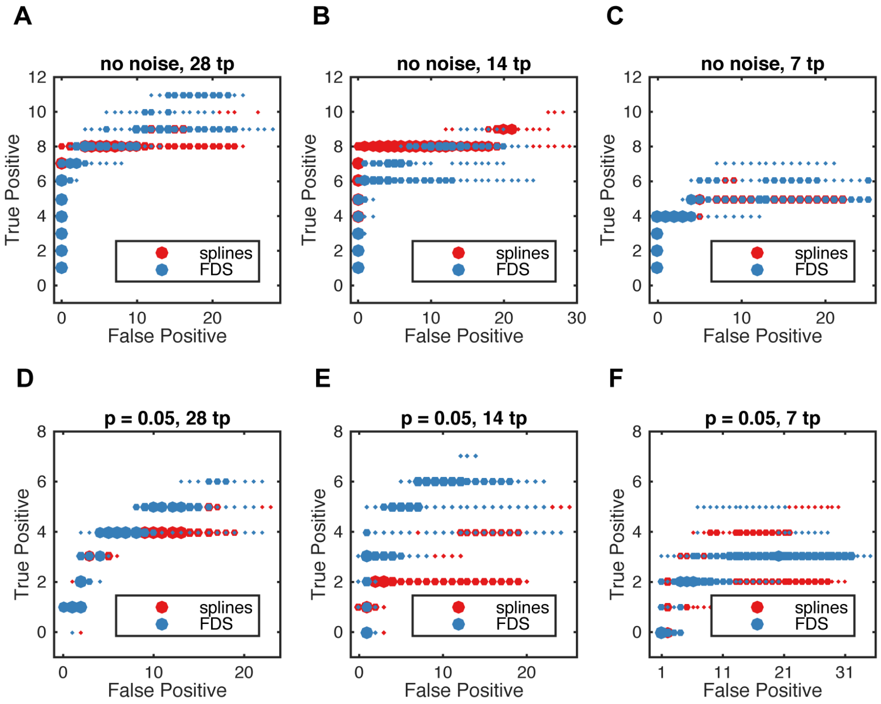


TruePos	FalsePos	TrueNeg	FalseNeg
5	1	2171	8

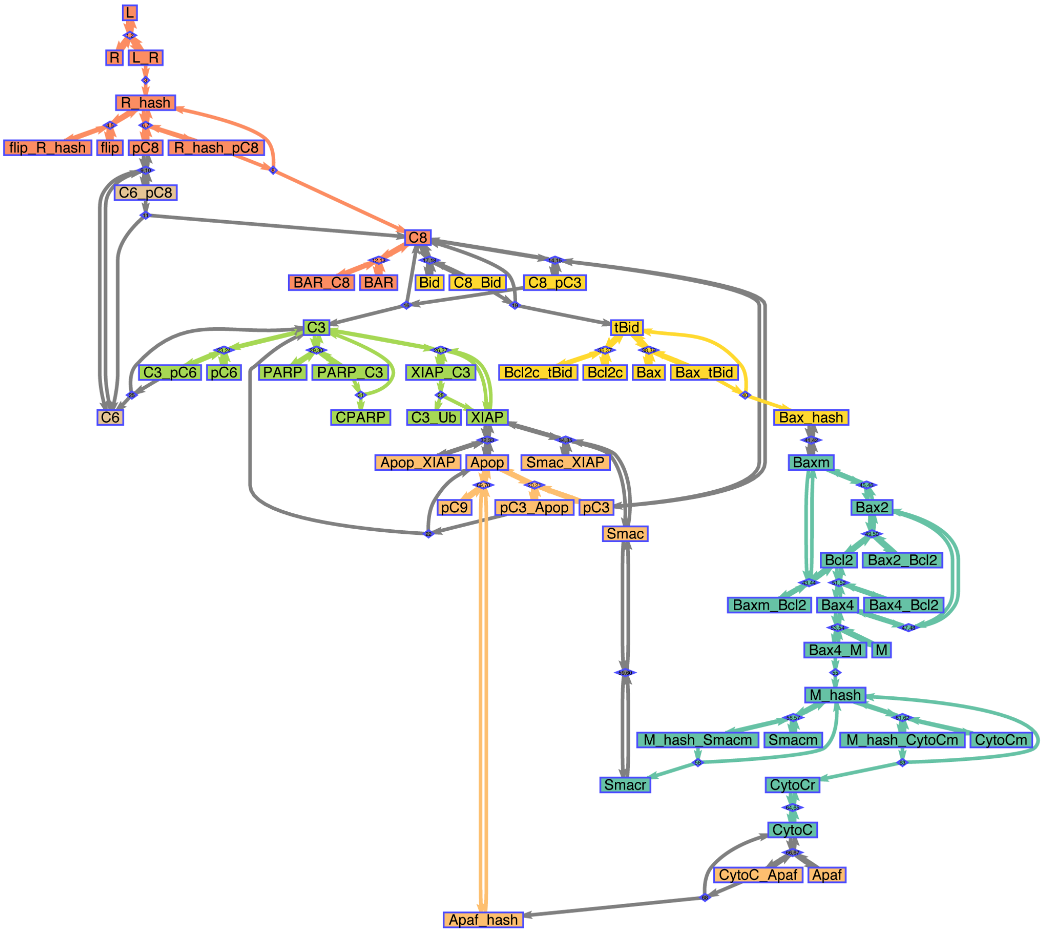
Supplementary Figure A.5.: Structure learning performance of the *reactionet lasso* for 10^5 single cell trajectories evaluated at 13 time points for apoptotic receptor subunit ($p = 0.05$). Empirical moment gradients estimated with cubic splines. Solution selected with Bayesian Information Criteria (BIC).



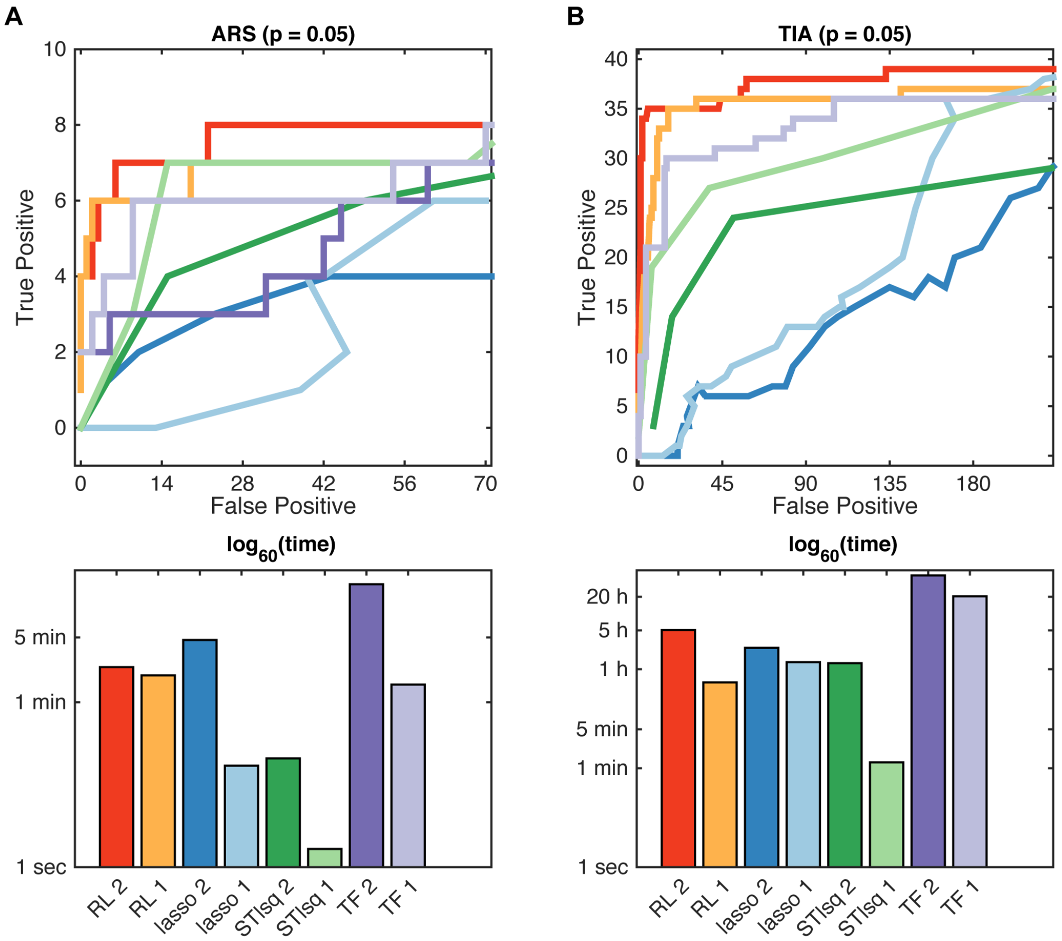
Supplementary Figure A.6.: Analysis of standard deviation of moment and stoichiometric moment function estimates for high order moments for different sample sizes. Results for application to apoptotic receptor subunit ($p = 0.05$). (A) Absolute values of standard deviation of moment estimate estimated from bootstrap for the apoptotic receptor subunit with no noise with 10^5 (red), 10^4 (blue), 10^3 (green) trajectories, 13 time points. (B) Relative change of standard deviation of the moment estimates with decreasing number of trajectories compared to 10^5. (C) Corresponding absolute and relative change of standard deviation of design matrix estimate (with stoichiometric moment functions as entries) with decreasing number of samples compared to 10^5.



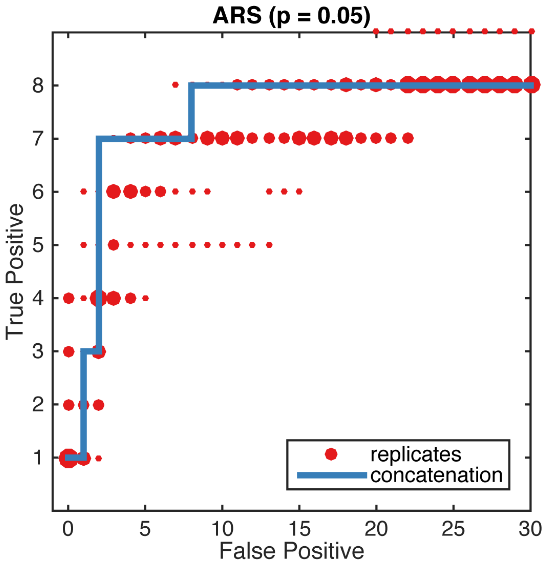
Supplementary Figure A.7.: Overlay of five regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit for uniform selection of timepoints. Results for different empirical moment gradient estimates: splines (red), FDS (blue) for different amount of time points and different levels of noise: 28 (A, D), 13 (B, E), 7 (C, F).


L
R
L_R
R_hash
flip_R_hash
flip
pC8
R_hash_pC8
C6_pC8
C8
BAR_C8
BAR
Bid
C8_Bid
C8_pC3
C3
tBid
C3_pC6
pC6
PARP
PARP_C3
XIAP_C3
Bcl2c_tBid
Bcl2c
Bax
Bax_tBid
C6
CPARP
C3_Ub
XIAP
Bax_hash
Apop_XIAP
Apop
Smac_XIAP
Baxm
pC9
pC3_Apop
pC3
Bax2
Smac
Bcl2
Bax2_Bcl2
Baxm_Bcl2
Bax4
Bax4_Bcl2
Bax4_M
M
M_hash
M_hash_Smacm
Smacm
M_hash_CytoCm
CytoCm
Smacr
CytoCr
CytoC
CytoC_Apaf
Apaf
Apaf_hash

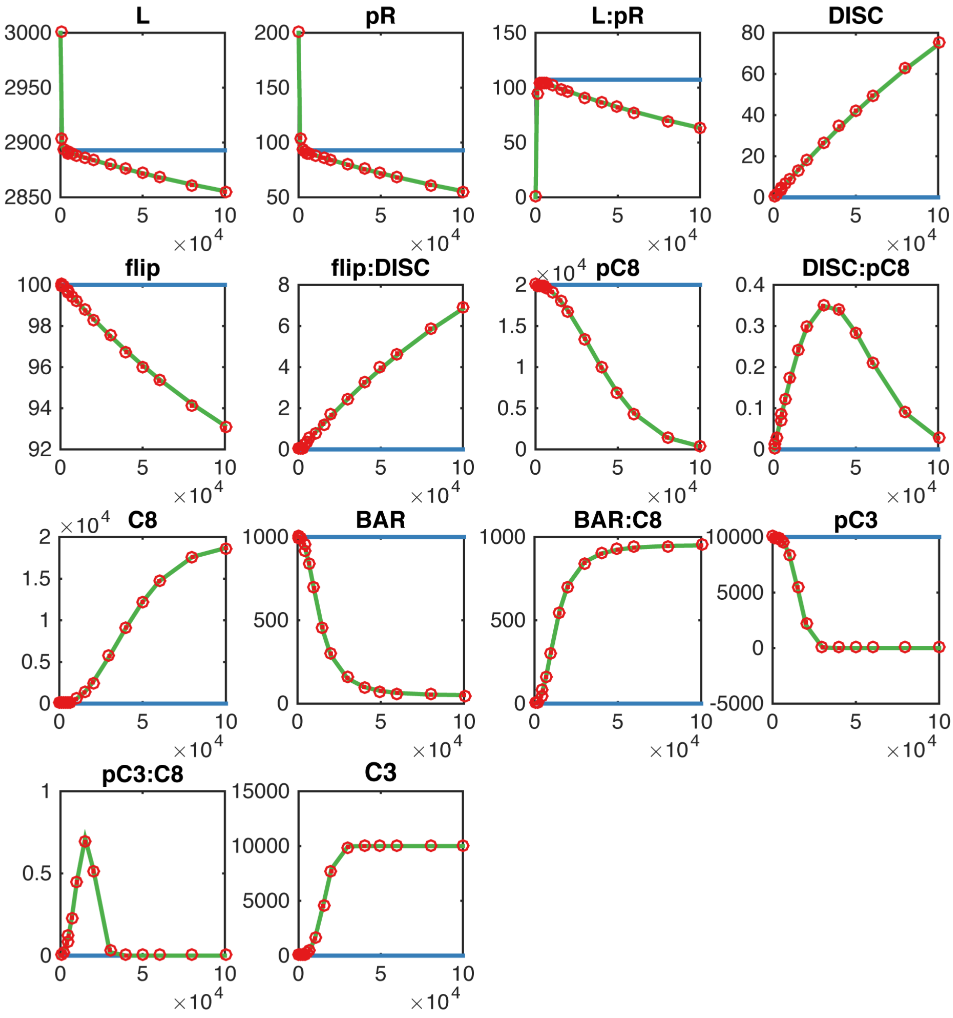
Supplementary Figure A.8.: Original reaction network of TRAIL induced apoptosis. Different modules colored in different colors. Reactions connecting the models depicted in gray.



Supplementary Figure A.9.: Comparison of the *reactionet lasso* with various simplified baseline procedures. RL = *reactionet lasso*; STlsq = sequential thresholded regression, TF = Topological filtering. All methods applied to Moment Equations of 1st and 2nd order correspondingly. Results for: **(A)** the apoptotic receptor subunit with noise (p = 0.05) with 10^5 trajectories, 13 time points; **(B)** TRAIL-induced apoptosis with noise (p = 0.05) with 10^5 trajectories, 33 time points. TF2 was interrupted after 2h hours and didn't produce any solution in the range of cardinality represented on the plot.

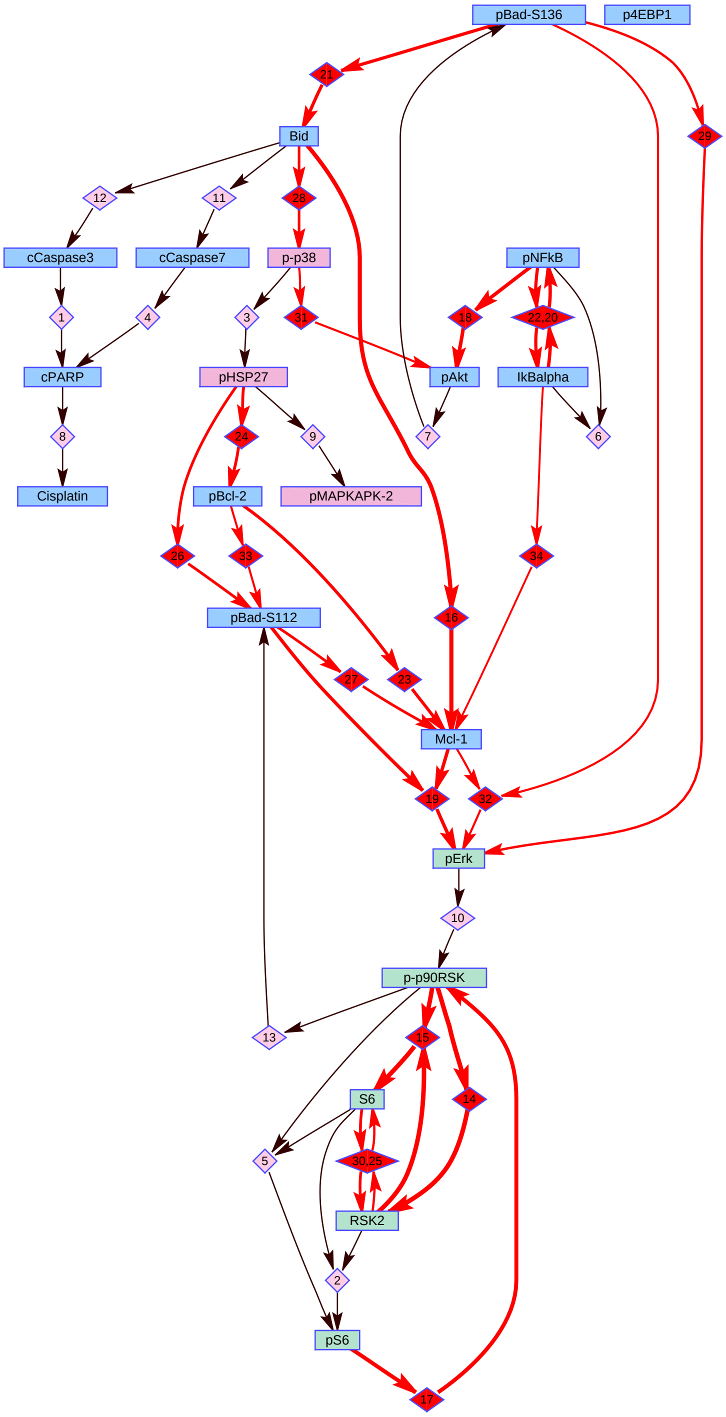


Supplementary Figure A.10.: Results for application of different strategies for *reactionet lasso* for the case of multiple replicates to apoptotic receptor subunit ($p = 0.05$). Red dots correspond to different replicates. Size of the dot proportional to the frequency of the solution between the replicates. Blue line corresponds to the strategy of concatenating design and response matrices.



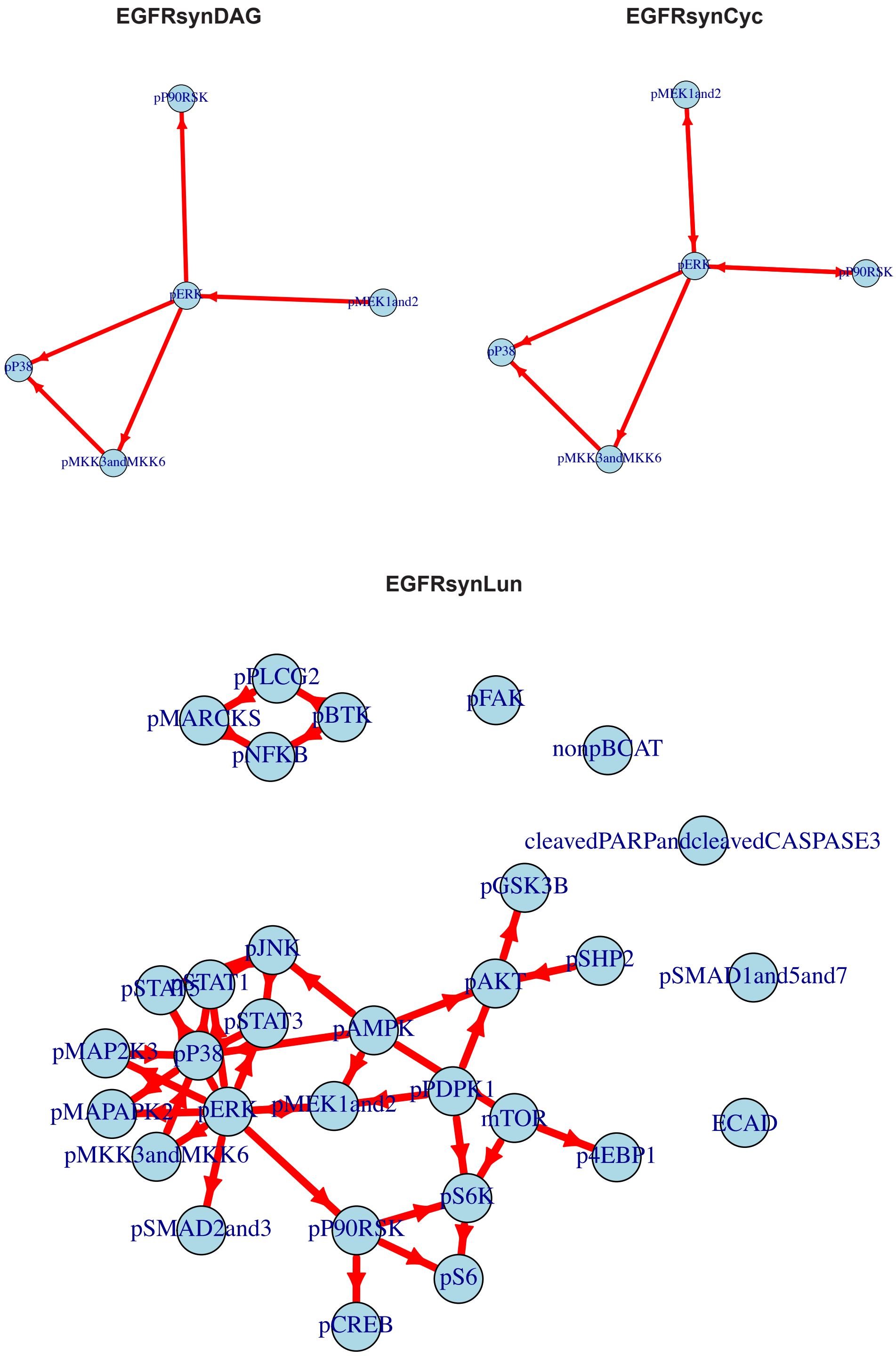
Supplementary Figure A.11.: Recovery of the dynamics of mean trajectories by the *reactionet lasso*. Red: observed data for 10^5 single cell trajectories evaluated at 13 time points for apoptotic receptor subunit without measurement noise. Solution selected with AIC for two distinct scenarios: *ab initio* learning (blue), *a priori* specified reaction identified false negative in *ab initio* learning setting (green).

B Supplementary material Chapter 3


pBad-S136
p4EBP1
21
29
Bid
12
11
28
cCaspase3
cCaspase7
p-p38
pNFkB
1
4
3
31
18
22,20
cPARP
pHSP27
pAkt
IkBalpha
8
24
9
7
6
Cisplatin
pBcl-2
pMAPKAPK-2
26
33
34
pBad-S112
16
27
23
Mcl-1
19
32
pErk
10
p-p90RSK
13
15
S6
14
5
30,25
RSK2
2
pS6
17

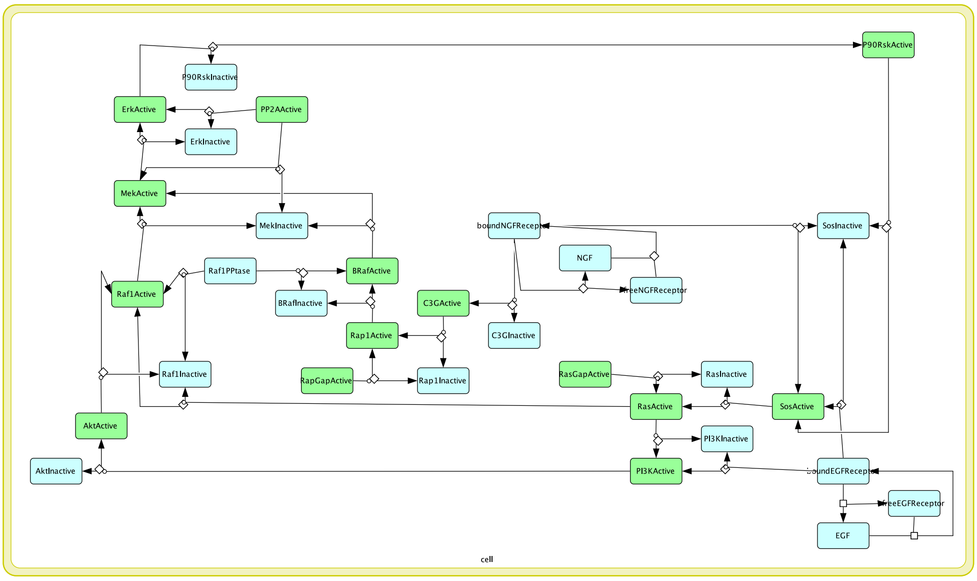
Supplementary Figure B.1.: Raw output of Reactionet lasso. Black edges correspond to the **fixed** reactions. The number assigned to a reaction corresponds to the relative confidence of this reaction: the lower the number, the more confident we are about the reaction.

C Supplementary material Chapter 4

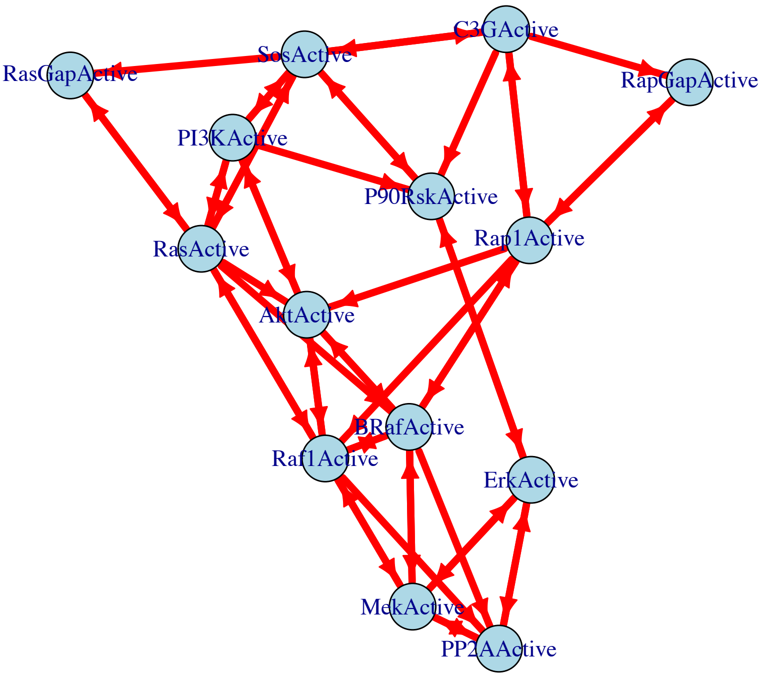

EGFRsynDAG
pP90RSK
pERK
pMEK1and2
pP38
pMKK3andMKK6
EGFRsynCyc
pMEK1and2
pERK
pP90RSK
pP38
pMKK3andMKK6
EGFRsynLun
pPLCG2
pMARCKS
pBTK
pNFKB
pFAK
nonpBCAT
cleavedPARPandcleavedCASPASE3
pGSK3B
pJNK
pSHP2
pSTAT5
pSTAT1
pAKT
pSMAD1and5and7
pSTAT3
pAMPK
pMAP2K3
pP38
pPDPK1
pMAPAPK2
pERK
pMEK1and2
mTOR
ECAD
pMKK3andMKK6
p4EBP1
pS6K
pSMAD2and3
pP90RSK
pS6
pCREB

Supplementary Figure C.1.: Causal graphs for synthetic models simulated from CTTSM.

SBML model


P90RskActive
P90RskInactive
ErkActive
PP2AActive
ErkInactive
MekActive
MekInactive
boundNGFRecept
SosInactive
NGF
freeNGFReceptor
Raf1PPtase
BRafActive
Raf1Active
BRafInactive
C3GActive
Rap1Active
C3GInactive
Raf1Inactive
RapGapActive
Rap1Inactive
RasGapActive
RasInactive
RasActive
SosActive
AktActive
PI3KInactive
AktInactive
PI3KActive
boundEGFRecept
freeEGFReceptor
EGF
cell

Causal graph


C3GActive
SosActive
RasGapActive
RapGapActive
PI3KActive
P90RskActive
RasActive
Rap1Active
AktActive
BRafActive
Raf1Active
ErkActive
MekActive
PP2AActive

Supplementary Figure C.2.: SBML model graph representing full mechanistic models (observed nodes are green) and the corresponding causal graph for a synthetic system generated from ODE model.

List of Figures

1.1. A simple taxonomy of models from Peters et al. in the light of modern Systems Biology (proteomics) research and contribution of this thesis. . 6

2.1. **Schematic representation of *reactionet lasso* procedure.** For details see section Methods. 18

2.2. **Performance assessment of the first *reactionet lasso* step for 10^5 single cell trajectories of the apoptotic receptor subunit.** (A) Enrichment and depletion of true and respectively false positive reactions for the reaction rate estimates $\hat{\mathbf{k}}^{LS}$ (red) and $\hat{\mathbf{k}}^{FG}$ (blue). Results are reported for gradient estimation procedures smooth, FDS, splines (see main text for details). (B) Comparison of response (empirical moment gradients) and prediction with feasible generalized least square estimate for moments of different order: means (blue), variances (green), covariances (yellow) and prediction with true rate constants for all moments (red crosses). 22

2.3. **Influence of measurement noise on structure learning capacity.** (A) Example of several single cell trajectories of one of the species (BAR) in apoptotic receptor subunit: without measurement noise (red), with measurement noise according to the binomial model with probability of success $p = 0.1$ (blue), 0.05 (green). Comparison of reconstructed means for known p between different noise levels shows how empirical moments are affected by measurement noise. Black dots represent snapshot measurements used for the inference procedure. (B) Overlay of five regularization paths in terms of true/false positive tradeoff for different measurement noise levels as indicated in the legend in terms of binomial capture efficiency. Structure learning performance for 10^5 single cell trajectories and thirteen time points of the apoptotic receptor subunit. Empirical moment gradients estimated with splines. 23

2.4. **Overlay of five regularization paths in terms of true/false positive tradeoff over different data availability situations.** Results for *reactionet lasso* application to apoptotic receptor subunit ($p = 0.05$). Empirical moment gradients estimated with cubic splines. (A) 10^5 single cell trajectories evaluated at different amount of time points: 28 (red), 13 (blue), 7 (green). (B) Different number of single cell trajectories: 10^5 (red), 10^4 (blue), 10^3 (green) evaluated at thirteen time points. 24

2.5. **Overlay of five regularization paths in terms of true/false positive tradeoff for different moment orders of gradients considered for structure learning:** up to 2nd order (red) Moment Equations, up to 1st order (blue) Moment Equations and only 1st order moments for deterministic mean ODE model (green). Structure learning performance for 10^5 single cell trajectories and thirteen time points of the apoptotic receptor subunit. Empirical moment gradients estimated with splines. Results represented for different levels of measurement noise: (A) no noise; (B) $p = 0.1$; (C) $p = 0.05$. 25

2.6. **Structure learning with prior knowledge for the 70 component TRAIL induced apoptosis cascade.** Structure learning performance for 10^5 single cell trajectories and 33 time points and capture efficiency $p = 0.05$. Empirical moment gradients estimated with splines. (A) Example of a recovered graph for the setting above and no prior knowledge. True positive reactions are red, false positive reactions are blue, false negative reactions are pink. (B) Regularization paths in terms of true/false positive tradeoff (including prior knowledge reactions, see **Results** for details) for different prior knowledge situations. Following prior knowledge situations are depicted: no prior knowledge (green). Additional prior knowledge situations comprise ten instances of 10%(blue)/ 50%(red) randomly chosen known reactions. Diameter of dots and color code indicate frequency of solutions with a specific true/false positive tradeoff. Black dots represent solutions coinciding with large improvements of BIC. 27

3.1. Mass cytometric profiling of TRAIL-treated HeLa cells reveals bimodal response across time. (A) HeLa cells treated with TRAIL collected at 30 min. intervals across six hours were analyzed by mass cytometry. Heatmap shows asinh ratio of median expression intensities of each protein marker (columns) across all time points measured (rows), relative to the untreated baseline (first row). (B) Histograms displaying signal intensities of representative markers displaying emergence of bimodality with exposure to TRAIL. 41

3.2. Identification of mechanism of fractional killing from single-cell data with optimal transport and Reactionet lasso. (A) Optimal transport matching for two consequent snapshots. Blue and grey dots correspond to “alive” and “dead” cells accordingly. (B) Time of death reconstruction from the transition matrix and death status: if alive cell transitioned to the dead state we assign the corresponding time of death. (C) “Rainbow” plots reconstruction: cells are split into different subpopulations by the time of death and the median expression of the markers in each subpopulation is depicted. (D) Entire analysis pipeline: hypothesis formulation with optimal transport and reactions refinement with Reactionet lasso. 44

3.3. Optimal transport matching identifies distinct trajectories of TRAIL-treated HeLa cells. The classification score shows steady increase in the classification accuracy between the apoptotic and survival populations. Feature score shows how significance of an individual feature changes over time. “Rainbow” plots show Asinh transformed median expression values of the most significant markers at the beginning of the process (t-0) in TRAIL-treated HeLa cells with different fates upon treatment with TRAIL. 46

3.4. “Rainbow” plots show Asinh transformed median expression values of some BCl-3 family markers, p-Erk and p-p38 in TRAIL-treated HeLa cells with various kinase inhibitors. 51

3.5. Single-cell time course data implicates selectively maintained IkBalpha activation and downstream MAPK/survival pathways in TRAIL-resistant subpopulation. (A) Simplified Reactionet lasso graph documenting signaling network relationships revealed during TRAIL response. Known relationships documented in literature and novel relationships discovered from mass cytometry time course data are shown in solid and dotted lines respectively. (B) Histogram summarizing mean percentage of surviving clones following treatment with TRAIL plus targeted kinase inhibitors relative to TRAIL plus DMSO vehicle control, as measured by clonogenic assay. 52
3.6. Antibodies used for mass cytometry experiments.. 56

4.1. General time series model for disjoint snapshots in the context of transfection experiments. **A** Time-unrolled model: individual box represents measured nodes per snapshot. We assume: (i) there are no instantaneous causal relations inside one snapshot, (ii) causal structure is preserved between pairs of consequent snapshots. Solid circles represent observed variables per snapshot. **B** Corresponding time-rolled causal model. Time-rolled graph is allowed to have cycles (negative and positive feedback loops). 65
4.2. *MassCaRA* pipeline. Green and blue boxes represent pooling of the data for different steps of the method. 67
4.3. Comparison of *MassCaRA* with different hyperparameteres against various settings of *ICP* and *backShift*. The comparison is done in terms of the ration of between true positives and false positives to the *total amount of positive edges*. We limited the x-axis to the value 1.0, because if the amount of false positives proposed by the method is greater than the total amount of positives, the results are not practical from biological point of view. 71
4.4. Output of *MassCaRA* on Lun et al. dataset. Only edges with score above 0.6 were depicted for readability. Red edges correspond to positive activation and blue ones to negative feedback. Directions of red and blue edges was identified by transient transfection and gray edges correspond to the ICP directions. 73

A.1. Structure learning performance of the *reactionet lasso* for 10^5 single cell trajectories evaluated at 13 time points for **(A)** apoptotic receptor subunit (no measurement noise); **(B)** the enzymatic system. Empirical moment gradients estimated with cubic splines. Solution selected with Bayesian Information Criteria (BIC). 100

A.2. Regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit (no measurement noise). **(A-B)** Empirical moment gradients estimated with "smooth" procedure: **(A)** 10^5 single cell trajectories evaluated at different amount of time points (tp) as indicated in the legend. **(B)** Different number of single cell trajectories: 10^3, 10^4, 10^5 evaluated at thirteen time points. **(C-E)** Results for different empirical moment gradient estimates: smooth(red), splines (blue), FDS (green) for different amount of time points: 28 (C), 13 (D), 7 (E). 101

A.3. Regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to enzymatic system (no measurement noise). **(A-B)** Empirical moment gradients estimated with "smooth" procedure: **(A)** 10^5 single cell trajectories evaluated at different amount of time points (tp) as indicated in the legend. **(B)** Different number of single cell trajectories: 10^3, 10^4, 10^5 evaluated at thirteen time points. **(C-E)** Results for different empirical moment gradient estimates: smooth(red), splines (blue), FDS (green) for different amount of time points: 28 (C), 13 (D), 7 (E). 102

A.4. Overlay of five regularization paths in terms of true/false positive tradeoff over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit (p = 0.05) with 13 time points. Results for different empirical moment gradient estimates: splines (red), FDS (blue) for different amount of time points: 28 (A), 13 (B), 7 (C). 103

A.5. Structure learning performance of the *reactionet lasso* for 10^5 single cell trajectories evaluated at 13 time points for apoptotic receptor subunit ($p = 0.05$). Empirical moment gradients estimated with cubic splines. Solution selected with Bayesian Information Criteria (BIC). 103

A.6. Analysis of standard deviation of moment and stoichiometric moment function estimates for high order moments for different sample sizes. Results for application to apoptotic receptor subunit (p = 0.05). (A) Absolute values of standard deviation of moment estimate estimated from bootstrap for the apoptotic receptor subunit with no noise with 10^5 (red), 10^4 (blue), 10^3 (green) trajectories, 13 time points. (B) Relative change of standard deviation of the moment estimates with decreasing number of trajectories compared to 10^5. (C) Corresponding absolute and relative change of standard deviation of design matrix estimate (with stoichiometric moment functions as entries) with decreasing number of samples compared to 10^5. 104

A.7. Overlay of five regularization paths in terms of true/false positive trade-off over different data availability situations. Results for *reactionet lasso* application to apoptotic receptor subunit for uniform selection of timepoints. Results for different empirical moment gradient estimates: splines (red), FDS (blue) for different amount of time points and different levels of noise: 28 (A, D), 13 (B, E), 7 (C, F). 105

A.8. Original reaction network of TRAIL induced apoptosis. Different modules colored in different colors. Reactions connecting the models depicted in gray. 106

A.9. Comparison of the *reactionet lasso* with various simplified baseline procedures. RL = *reactionet lasso*; STlsq = sequential thresholded regression, TF = Topological filtering. All methods applied to Moment Equations of 1st and 2nd order correspondingly. Results for: **(A)** the apoptotic receptor subunit with noise (p = 0.05) with 10^5 trajectories, 13 time points; **(B)** TRAIL-induced apoptosis with noise (p = 0.05) with 10^5 trajectories, 33 time points. TF2 was interrupted after 2h hours and didn't produce any solution in the range of cardinality represented on the plot. 107

A.10. Results for application of different strategies for *reactionet lasso* for the case of multiple replicates to apoptotic receptor subunit ($p = 0.05$). Red dots correspond to different replicates. Size of the dot proportional to the frequency of the solution between the replicates. Blue line corresponds to the strategy of concatenating design and response matrices. 108

A.11. Recovery of the dynamics of mean trajectories by the *reactionet lasso*. Red: observed data for 10^5 single cell trajectories evaluated at 13 time points for apoptotic receptor subunit without measurement noise. Solution selected with AIC for two distinct scenarios: *ab initio* learning (blue), *a priori* specified reaction identified false negative in *ab initio* learning setting (green). 109

B.1. Raw output of Reactionet lasso. Black edges correspond to the **fixed** reactions. The number assigned to a reaction corresponds to the relative confidence of this reaction: the lower the number, the more confident we are about the reaction. 112

C.1. Causal graphs for synthetic models simulated from CTTSM. 114
C.2. SBML model graph representing full mechanistic models (observed nodes are green) and the corresponding causal graph for a synthetic system generated from ODE model. 115

7 Curriculum vitae

Anna Klimovskaia

Curriculum Vitae

Education

July 2013 – present	**PhD student, Research Assistant**, *ETH Zurich, Institute of Systems Molecular Biology*, Zurich.
June 2013	**Specialist in Mathematics and System Programming (Master equivalent)**, *Lomonosov Moscow State University*, Diploma with Honours, GPA: *4.9/5*. The Faculty of Computational Mathematics and Cybernetics, Chair of Mathematical Statistics, Laboratory of Statistical Analysis.
June 2007	**High School Diploma**, *Advanced Educational Scientific Center of Lomonosov Moscow State University — Kolmogorov School*. Advanced mathematics class

Additional

July 2018	*PAISS, Artificial Intelligence Summer School*, project.inria.fr
August 2017	*Data Science Summer School*, www.ds3-datascience-polytechnique.fr
July 2017	*Bocconi Summer School in Advanced Statistics and Probability: Statistical and Causal Learning*, spas.lakecomoschool.org
July 2006	*Summer School in Computer Science*, www.olympiads.ru

Awards

2017 academic year	BIDSA (Bocconi Institute of Data Science) award for an outstanding applicant for Bocconi Summer School in Advanced Statistics and Probability
2011/2012 academic year	Russian President scholarship for talented students
Spring 2010	Patent 2011610590 from 11.01.2011. Software package for computation of probability distribution, density and moments for Bayesian exponential-Erlang models of information flows. Authors: S. Shorgin, A. Klimovskaya, A. Kudryavtsev
2007	III diploma Moscow high school olympiad in informatics
2007	III diploma Moscow high school programming contest (ACM ICPC Subregional Contest equivalent for high school students)

Experience

January 2018 – May 2018	**Research Intern**, *Facebook Artificial Intelligence Research (FAIR)*, New York, Supervisor: Leon Bottou, Maximilian Nickel.

June 2011 – Jan 2013	**Junior Analyst**, *Yandex, Group of Search Analysis*, Moscow.
Responsibilities	Research work under minimum supervision; Statistical description of consumer behaviour; Extracting information from search logs on Yandex's servers using MapReduce; Analysis of large amounts of data; Data quality management; Submitting written reports on every task

Research projects

Doctoral Thesis

2013-now	*Identification of causal mechanisms of cell variabilty in signaling networks from single-cell time series snapshots*
Supervisors	Manfred Claassen, Assistant Professor, PhD
Description	Development of algorithms to learn causal interactions from single-cell data.

Master Thesis

2011-2013	*Analysis of probability distributions arising in high-frequency trading*
Supervisors	Doynikov Alexander, Research Fellow, PhD
Description	Time series analysis on the high-frequency data on futures S&P 500

Applied research project

2011-2012	*Automatic bursts and changes of trend detection in large-scaled multidimensional non-stationary time series.*
Company	Yandex (Russian largest Internet search engine)
Description	The goal was to create a tool for automatic monitoring of different indicators of market share. I introduced a stochastic model and developed a self-learning algorithm based on it

Course Work

2010-2011	*Computation of implied volatility*
Supervisors	Doynikov Alexander, Research Fellow, PhD
Description	Analysis of different approaches for computation of implied volatility: Newton-Raphson method, Tailor series expansion and rational approximation. The methods were implemented in a C++ code. The accuracy and performance comparison of the methods was made

Other Projects

Spring 2010	*Research project on bayesian approach to the analysis of queueing systems.*
Supervisor	Kudryavtsev Alexey, Senior Lecturer, PhD
Description	My main role was the development of software package for computation of probability distribution, density and moments for Bayesian exponential-Erlang models of information flows. A patent was acquired (see Awards)

Publications

December 2016	**Klimovskaia A**, Ganscha S, Claassen M (2016) Sparse Regression Based Structure Learning of Stochastic Reaction Networks from Single Cell Snapshot Time Series. PLoS Comput Biol 12(12): e1005234. doi:10.1371/journal.pcbi.1005234
October 2018	Blom T, **Klimovskaia A**, Magliacane S, Mooij J M (2018) An Upper Bound for Random Measurement Error in Causal Discovery. ParXiv preprint arXiv:1810.07973

Languages

Russian	**Native**
English	**Fluent**
German	**Basic**

Computer skills

Application software	MATLAB, R; LaTex; MS Office: MS Excel (+ VBA, Evrista), MS Word, MS PowerPoint, MS Access;
Software development tools	Python (+ PyTorch, MapReduce); C++, C;

Qualities

Personal skills:	Strong analytical, problem-solving and communication skills. Good project management and presentation skills.