12-2016

Bayesian causal inference of cell signal transduction from proteomics experiments

Robert D. O. Ness
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Robert Ness

Entitled
Bayesian Causal Inference of Cell Signal Transduction from Proteomics Experiments

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Hyonho Chun
Chair

Bruce Craig

Michael Zhu

Rebecca Doerge

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Hyonho Chun

Approved by: Jun Xie 12/5/2016
Head of the Departmental Graduate Program Date

BAYESIAN CAUSAL INFERENCE OF CELL SIGNAL TRANSDUCTION FROM PROTEOMICS EXPERIMENTS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Robert D. O. Ness III

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

I dedicate this dissertation to my wife Shan. Without her support, time, labor, and bottomless well of patience, this would not have happened.

ACKNOWLEDGMENTS

I owe a debt of gratitude to my advisor, Dr. Olga Vitek, for her mentorship and financial support, for her encouragement and generosity with her time, and especially her vision, which has influenced me deeply as a statistician. She taught me how to get lab scientists excited about statistics, through effectively communicating with collaborators, as well as writing and presenting impactful conference and journal papers to an audience with varying backgrounds.

I also owe a great deal to my collaborator, mentor, and friend Dr. Karen Sachs. It was a tremendous opportunity just to learn science from her and collaborate with her on much of this work. Beyond that, she taught me how to be a graduate student, how to be passionate about the work, and how to enjoy myself. I am a better statistician and person for having known her.

Special and sincere thanks go to my committee members, Dr. Bruce Craig, Dr. Rebecca Doerge, Dr. Hyonho Chun, and Dr. Michael Yu Zhu for their very helpful comments and suggestions on my thesis work. I further express my gratitude to Dr. Doerge, for granting me four years of financial support through the Purdue Fellowship and offering essential advice on my Ph.D. study and research, and to Dr. Chun for sparking my interest in computational systems biology. Without their intervention I would not have been able to complete this dissertation.

I thank Dr. Parag Mallick for teaching me the nuts and bolts of systems biology, making crucial introductions within the systems biology community, and for providing financial support. Many thanks to the Purdue Department of Statistics, especially for their flexibility in allowing me to do research *in absentia*; in particular, I want to thank Dr. Doerge, Dr. Jun Xia, Dr. Chun, Ms. Anna Hook, and Ms. Marian Cannova in this regard. Thank you to Dr. Ahmed Elmagarmid and Dr. Halima Bensmail of QCRI for funding my research and hosting me in Doha. Thank you to

Dr. Marina Bessarabova and Dr. Yuri Nikolsky for hosting me as an intern in the computational biology group at Thomson Reuters. I am also very grateful to Dr. William Cleveland and Dr. Thomas Kuczeketc, for their support in my Ph.D study. I hope to thank Mr. Douglas Crabill and David LeFevre for their support in running simulations in the cluster. Finally, I thank Ms. Marian Cannova, Ms. Marian Duncan, Ms. Hook, Ms. Marshay Jolly, Ms. Diane Martin, Ms. Becca Pillion, Ms. Shaun Ponder and Ms. Mary Roe for their administrative support.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABBREVIATIONS

DAG	directed acyclic graph
CPD	conditional probability distribution
PDAG	partially-directed acyclic graph

ABSTRACT

Ness, Robert D. O. PhD, Purdue University, December 2016. Bayesian Causal Inference of Cell Signal Transduction from Proteomics Experiments . Major Professor: Hyonho Chun.

Cell signal transduction describes how a cell senses and processes signals from the environment using networks of interacting proteins. In computational systems biology, investigators apply machine learning methods for causal inference to develop causal Bayesian network models of signal transduction from experimental data. Directed edges in the network represent causal regulatory relationships, and the model can be used to predict the effects of interventions to signal transduction. Causal inference approaches applied to proteomics experiments use statistical associations between observed signaling protein concentrations to infer a causal Bayesian network model, but there is no experimental and analysis framework for applying these methods to this experimental context.

The goal of this dissertation is to provide a Bayesian experimental design and modeling framework for causal inference of signal transduction. We evaluate how different high-throughput experimental settings affect the performance of algorithms that detect conditional dependence relationships between proteins. We present a Bayesian active learning approach for designing intervention experiments that reveal the direction of causal influence between proteins. Finally, we present a Bayesian model for inferring the parameters of the conditional probability density functions in a causal Bayesian network. The parameters are directly interpretable as a function of the rate constants in the biochemical reactions between interacting proteins. The work pays special attention to analysis of single-cell "snapshot" data such as mass cytometry, where each cell is a multivariate cell-level replicate of signal trans-

duction at a single time point. We also address the role of large-scale bulk experiments such as mass-spectrometry-based proteomics, and small-scale time-course experiments in causal inference.

1. INTRODUCTION

Cell signal transduction describes how a cell senses and processes signals from the environment using networks of interacting proteins [1]. Many disease phenotypes stem from errors in signal transduction, including some cancers and autoimmune diseases. This dissertation addresses the problem of building a causal Bayesian network model of a signal transduction network from proteomics data. The benefits of this modeling framework are that its directed acyclic graph (DAG) structure represents the causal (regulatory) relationships between signaling proteins. Further, investigators can use the model to predict the effects of interventions in signaling (e.g. drug interventions), facilitating the development of new treatments.

1.1 Statement of biological problem

Causal inference in the context of signal transduction means determining the regulatory relationships between proteins. High-throughput proteomics experiments quantify the activity of signaling proteins. There are 3 tasks in causal inference from these experiments: (1) determine the presence of regulatory relationships; (2) determine the direction of a regulatory relationship given its presence; and (3) determine the magnitude of a regulatory effect given the relationship's presence and direction. Each task has different informational requirements from the experimental design. We describe each task and related experimental considerations as follows.

Task 1: Determine the presence of regulatory relationships. High-throughput proteomics platforms vary in the number of biomolecular analytes they can simultaneously quantify. For example, targeted mass spectrometry can quantify thousands of proteins in a sample, while the number of proteins quantified by capture sandwich immunoassays is lower by at least one order of magnitude. However, capture

immunoassay experiments have much higher sampling-throughput than targeted mass spectrometry [2]. Other dimensions in which the platforms differ include the precision of the quantifications, the ability to incorporate prior knowledge, and the ability to quantify proteins at single cell-level resolution. Experimentalists lack a design framework for matching the capabilities of different platforms to the data needs of statistical approaches for detecting regulatory relationships. This dissertation provides guidelines for selecting the right platform and experimental settings for this task.

Task 2: Determine the direction of a regulatory relationship given its presence. Given knowledge of a regulatory relationship between two signaling proteins, there are several sources of information on direction, *i.e.*, which protein is the regulator and which is the target. Pathway databases such as KEGG [20] contain canonical signaling pathways, constructed from curation of peer-reviewed literature. However, canonical pathways may not address signaling in the specific environmental conditions of interest to the investigator. Investigators may also have internal information from experiments conducted in the past on the signaling system of interest. However, these *historic datasets* also may have been collected under different experimental conditions. It is not clear how to make use of information from these sources in light of these incongruities. This dissertation provides a Bayesian approach to structuring prior knowledge from pathway databases and historic datasets.

In cell signaling studies, various perturbations are applied to samples across experimental conditions. Perturbations introduce variation in the signaling response [2, 3], providing more information to the statistical analysis. *Targeted interventions*, such as small molecule inhibitors, are a specific type of perturbation whose primary function is to determine the direction of a regulatory relationship [4–6]. However, targeted interventions add to the cost and complexity of an experiment. This dis-

sertation provides a cost-conscious experimental design framework for selecting targeted interventions.

Task 3: Determine the magnitude of regulatory effect given the relationship's presence and direction. To say that one signaling protein "regulates" another, means that these two proteins interact in one or more biochemical reactions. The magnitude of the regulatory effect depends on the rates at which these biochemical reactions occur. Reaction rates are determined by the abundance of the interacting signaling proteins, and a set of rate parameters [7]. Rate parameters describe the change in abundance of proteins in a signaling pathway in time. A quantification of the regulatory effect of one protein on another in terms of anything but rate parameters lacks biological interpretation. However, the technologies discussed so far do "snapshot" proteomics; they only quantify signaling at one single time point from the entire time course of the system's evolution. In the case of single cell proteomics technologies, this can be hundreds of thousands or even millions of cell-level snapshots. But even in this case, it is not clear what snapshots can reveal about rate parameters, a dynamic attribute of the system.

1.2 Statement of statistical problem

This dissertation addresses the objective of inferring a causal Bayesian network from proteomic studies of cell signaling. We rephrase the above three tasks as statistical inference tasks, each providing deeper insight into the signaling mechanism than the previous task: (1) infer the presence of causal edges; (2) infer the direction of the causal edge conditional on its presence; and (3) infer the magnitude of causal influence conditional on the presence and direction of the edge. We detail each of these inference tasks below.

Task 1: Infer the presence of causal edges. The first step of causal inference is to identify undirected edges. The structure of an undirected graphical model, or *Markov network*, represents conditional independence in the joint distribution over

all variables in the network [5]. *Conditional independence algorithms* search for statistical evidence of conditional independence in multivariate data [5, 27, 30], producing putative undirected edges. In some cases, and under key assumptions (such as no hidden confounders, see [10]) the undirected edges are causal relationships where the direction of causality is unknown. In protein signaling studies, these causal edges represent regulatory relationships between proteins. Applying conditional independence algorithms to proteomics experimental data is therefore a means of generating hypotheses of regulatory relationships [8–11]. This undirected edge hypothesis can be tested in a validation experiment targeting the two proteins that share the edge.

In high-throughput proteomics investigations, the relationship between the varying attributes of proteomics experimental platforms (e.g. protein-coverage and feasible sample size) and the results provided by conditional independence algorithms, has not been examined. Experimentalists know generally that increasing sample size improves the results of statistical analysis, and that false positives are a challenge in high-throughput experiments. However, the relationship between their choice of proteomics platform and the performance in edge detection is unclear. For example, if quantifying thousands of proteins in a targeted mass spectrometry experiment, it is not clear if using a sample size deemed larger than average for that platform would improve detection results by a meaningful degree. Nor is it clear if even better performance would be achieved with an assay that only targets hundreds of proteins (meaning lost opportunities for discovery) but with far greater sampling throughput. Further, experimentalists have the option of targeting proteins they know are more likely to regulate one another. It is unclear to experimentalists if this prior knowledge can alleviate problems with algorithm performance in their platform of choice. This dissertation uses a simulation analysis to demonstrate, in terms of sensitivity and specificity, how the dimension of the data and amount of

true interactions in the data affect the detection of undirected edges with conditional independence algorithms.

Task 2: Infer the direction of the causal edge conditional on its presence. In causal inference, targeted interventions fix the state of a random variable, revealing edge direction (causality) between it and other variables with whom it shares an edge [4–6]. Statistical association can be used to infer the presence of an undirected edge, but generally it is not sufficient for inference of edge direction. Targeted interventions are necessary to fully resolve the direction of causality [4, 5, 10]. In cell signaling studies, the experimental design for causal inference includes several elements: (1) an assay targeting the signaling proteins with causal interactions; (2) perturbation conditions that activate and vary the signaling response; (3) sample size sufficient for inference of DAG structure; and (4) targeted interventions. Due to the costs of proteomics experiments, an experiment typically includes a *batch* of targeted interventions (as opposed to one intervention per experiment). In theory for an undirected graph with p nodes, a batch of at most $p - 1$ targeted interventions is needed to fully resolve causality [12]. In practice, many of the $p - 1$ interventions would be redundant, and therefore would add to the cost and complexity of a causal inference experiment without contributing to causal inference. *Active learning* describes the machine learning task of optimal selection of targeted interventions [12–18]. In the context of causal inference in proteomics, existing methods typically assume one intervention is applied per experiment. This dissertation proposes an active learning method for selection of a batch of between 0 and $p - 1$ targeted interventions; enough to maximize causal information, while few enough that experimental cost and complexity are manageable.

The presence of an edge is inferred as well as the edge's direction, so an active learning approach must allow for uncertainty in as yet undirected edges; it is wasteful to use a targeted intervention to orient a false positive edge. Bayesian approaches to inferring DAG structure can address this uncertainty. However, these require constructing a prior distribution on the space of DAG structures. Current methods for

doing so require building a prior based on the ordering of nodes in the DAG [54]. The prior knowledge biologists have about signaling networks are not easily translated into DAG orderings.

Bayesian approaches address uncertainty in edge direction with *equivalence classes* of DAGs. An equivalence class is a set of DAGs with the same conditional independence structure (*i.e.*, same edge skeleton) and same posterior probability, but with different edge directions. Given a DAG structure inferred by causal inference, a different DAG from the same equivalence class is an equally probable causal explanation of the data. In graphical modeling, *DAG-to-PDAG algorithms* convert a DAG to a *partially-directed acyclic graph* (PDAG) [17, 55]. The PDAG represents the input DAG's equivalence class. The PDAG contains directed and undirected edges, the undirected edges correspond to edges with conflicting direction amongst members of the class. However, existing DAG-to-PDAG algorithms are not compatible with Bayesian causal inference. They will not produce the correct equivalence class if the DAG prior encodes causal information.

This dissertation proposes a method of incorporating prior causal information into a prior distribution on DAG structures that works with the tools biologists are familiar with. It also provides a DAG-to-PDAG algorithm that works with this prior distribution.

Task 3: Estimating causal influence and rate parameters. Given the presence and direction of an edge in the DAG, the causal Bayesian network model quantifies the strength of causal influence with a conditional probability distribution (CPD) [10]. The conditional probability of the activity of the protein (the effect) given the activity a direct regulator protein (the cause) quantifies both the magnitude and certainty of the cause-effect relationship. However, in prior work investigators' choice of the type of conditional probability distribution, such as Gaussian or multinomial, is governed primarily by practical convenience [11, 19]. These choices either fail to capture nonlinearity in signaling relationships (as with the

Gaussian) or are parameterized in a way that has no connection to the biology (as with the multinomial).

If we view conditional probability as a stochastic function that relates cause to effect, then it must have some relation to biochemical reaction rate parameters; causal influence is logically greater in the case where a regulator reacts with its target very often, than in the case where the reaction occurs rarely. Estimation of the magnitude of causal influence, and of the rate parameters themselves, remains an open problem in the causal inference of cell signaling with proteomics experiments. Further, when feedback loops are present in the set of biochemical reactions, it is not clear how to model the system with a causal Bayesian network at all, due to the acyclicity constraint in the DAG. This dissertation provides a conditional probability model parameterized in terms of rate parameters, and addresses the problem of feedback loops.

1.3 Contributions

- **Simulation study of edge detection supports task 1.** This dissertation presents simulations that interrogated the effectiveness of conditional independence detection algorithms in the high dimension, low sample-size settings common in high-throughput proteomics experiments. The analysis examines dimensions and sample-sizes that align with the coverage capabilities and sampling-throughput of proteomics platforms used in cell signaling studies. In addition, the analysis also considers the ability for the experimentalist to target proteins that are known to have a higher amount of interactions. The analysis evaluates sensitivity and specificity in edge detection, and provides guidance for a sequential experimental design in terms of these performance measures.

- **Causal prior DAG distribution supports task 2.** This dissertation proposes a method for building a prior distribution on the space of DAGs using

canonical pathway databases and historic data. The approach allows the experimentalist to encode canonical causal knowledge, in the form of the presence and orientation of an edge in a canonical database, into the prior. This provides a simpler interface than other Bayesian methods for DAG inference because biologists are already familiar with pathway representations of signaling. Historic data is taken from the investigator's own past experiments or from public repositories, and is not required to have come from an experimental design specifically targeting causal inference. This prior distribution encodes all available public and internal causal information about the signaling system under study prior to conducting a causal inference experiment.

- **DAG-to-PDAG algorithm that accounts for causal prior supports task 2.** The proposed prior is not compatible with causal inference methods that work with DAGs and equivalence classes. DAG-to-PDAG algorithms do not produce a PDAG representing posterior-equivalent DAGs if the prior contains causal information. This dissertation proposes an algorithm that corrects this, *i.e.*, it takes DAG and information about the prior as arguments and produces a PDAG representing a posterior-equivalence class. Beyond systems biology, this algorithm can be applied in causal inference problems where there is prior knowledge on causality.

- **Bayesian active learning algorithm supports task 2.** This dissertation provides a Bayesian active learning method for selecting targeted interventions. The approach quantifies the amount of causal uncertainty present in a probability distribution of DAGs, and derives a metric that quantifies how much and a given intervention would reduce uncertainty in the distribution. The metric is derived such that its calculation is easily parallelized for faster computing times.

- **Stopping criteria for selecting interventions supports task 2.** The work proposes a Bayesian stopping criteria for adding to the batch of inter-

ventions in an experiment. The active learning approach iteratively adds to the batch the intervention that has the highest expected contribution to causal inference from all candidate interventions. The stopping rule is triggered when that expected contribution falls below a certain threshold. This prevents the wasteful application of interventions in a causal inference experiment.

- **Probability model of steady-state signaling supports task 3.** We introduce an experimental constraint for single cell proteomics experiments, wherein protein abundance is quantified after the signaling system has reached a stable steady state. We show if the stable steady state assumption can be applied, we can model signaling regulation with the stationary distribution of a Markov process describing the statistical mechanics of biochemical reactions underlying signaling. We use the stationary distribution for each protein as the conditional probability distribution for the corresponding node in a DAG. This produces a causal Bayesian network with two attractive features: (1) the conditional probability distributions are parameterized in terms of rate parameters, and (2) we explicitly model biological stochasticity and connect it to cell-to-cell variation in single cell data. Further, the constraint allows us to model biochemical feedback loops with a causal Bayesian network model, without violating the acyclicity constraint. We provide a Bayesian modeling framework for inferring the parameters, and provide an algorithm that generalizes the code to any DAG structure, avoiding the need to code a new model for each new experiment.

2. STATISTICAL APPROACHES TO LEARNING REGULATORY RELATIONSHIPS IN LARGE-SCALE BIOMOLECULAR INVESTIGATIONS

2.1 Introduction

Modern high-throughput technologies such as mass spectrometry simultaneously quantify hundreds or even thousands of biomolecular analytes. Statistical associations (e.g. Pearson correlation, Spearman correlation, and mutual information) between observed protein concentrations suggest an enticing number of hypotheses regarding the underlying causal biomolecular mechanisms. However, associations do not imply causation. Formal methods of causal inference [9, 10] are required to probe these statistical associations for causality, and infer the underlying regulatory mechanisms.

Causal inference is increasingly of interest in proteome research. It has previously been used to learn the directed structure of signal transduction networks, e.g. in reconstruction of the T-cell signaling network from flow cytometry investigations in Sachs et al. 2005 [21], and the human liver carcinoma cell signaling network from Luminex antibody/bead-based XMAP technology in Saez-Rodriguez et al. 2009 [22] and participants in the DREAM4 Predictive Signaling Network Challenge [23]. However, the successful examples of causal inference in proteomics are very few [24]. In this perspective we argue that this is not an accident. The difficulty stems from the fact that large numbers of quantified analytes, combined with small sample size, lead to more spurious pairwise associations, obfuscate the true signal, and increase the false discoveries of putative causal events. Below we describe in non-technical terms the process of elucidating causal associations from high-throughput data, and suggest practical approaches for causal inference in large

scale proteomic datasets. Specifically, we suggest that the task of causal inference can be facilitated by refining the biological question, and by improving experimental design in terms of selection of (1) the subset of analytes, (2) the number of biological replicates, and (3) the type of biological conditions and stresses.

2.2 Background

2.2.1 Small-scale statistical inference of causal relationships: conditional independence and interventions

Consider, e.g. the MAPK signaling cascade in Figure 2.1, which is part of several signaling pathways such as the EGFR MAPK pathway [25]. In this cascade Raf causally affects the level of active (i.e., phosphorylated) Mek, while Mek causally affects Erk. Imagine these causal relationships were unknown: could they be detected from quantitative measurements on these phosphoproteins?

To illustrate the process of causal inference we simulated artificial data using the computational Huang-Ferrell model [26] of this cascade. The model represents the key binding, phosphorylation, and dephosphorylation reactions of the cascade with mass action kinetics, and replicates the MAPK key signaling behavior observed in nature. We used the model to simulate an experiment with 50 replicate biological samples, and measurements of concentration (μmol) of phosphorylated Raf, and doubly phosphorylated Mek and Erk in each sample.

Figure 2.2 (a) demonstrates the causal inference workflow starting with analysis of statistical associations in the data. In step 1, a correlation graph between cascade components Raf, Mek, and Erk is assembled from the measurements of protein concentration. Step 2 reduces the correlation graph to a sparse graph of inferred underlying conditional dependencies (Raf–Mek, and Mek–Erk). Step 3 interrogates this graph to find putative directions of causal relationships (Raf $\rightarrow$ Mek, and Mek $\rightarrow$ Erk). While step 1 has little requirements, step 2 requires that the number of

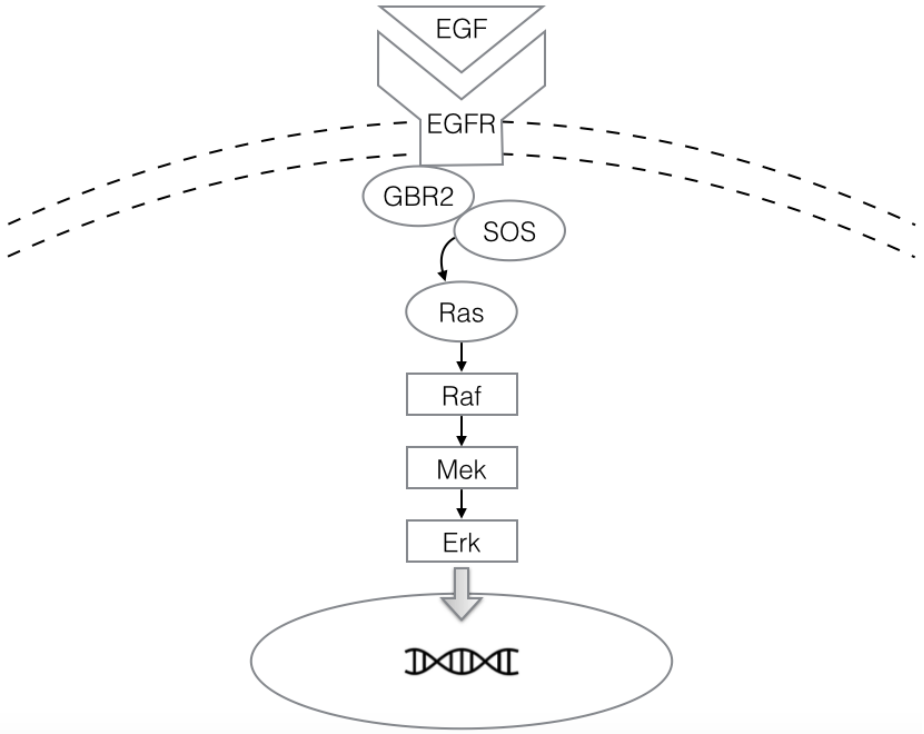

Fig. 2.1. EGFR MAPK signaling pathway, an example of a pathway containing the phosphorylation cascade from Raf to Mek to Erk. The binding of ligand EGF to EGFR initiates a signal that leads to the cascade, which in turn regulates transcription. This cascade implies two direct causal relationships, namely Raf $\rightarrow$ MEK, and Mek $\rightarrow$ Erk. Raf and Erk have an indirect causal relationship, through Mek. Figure republished with permission from original source [46]

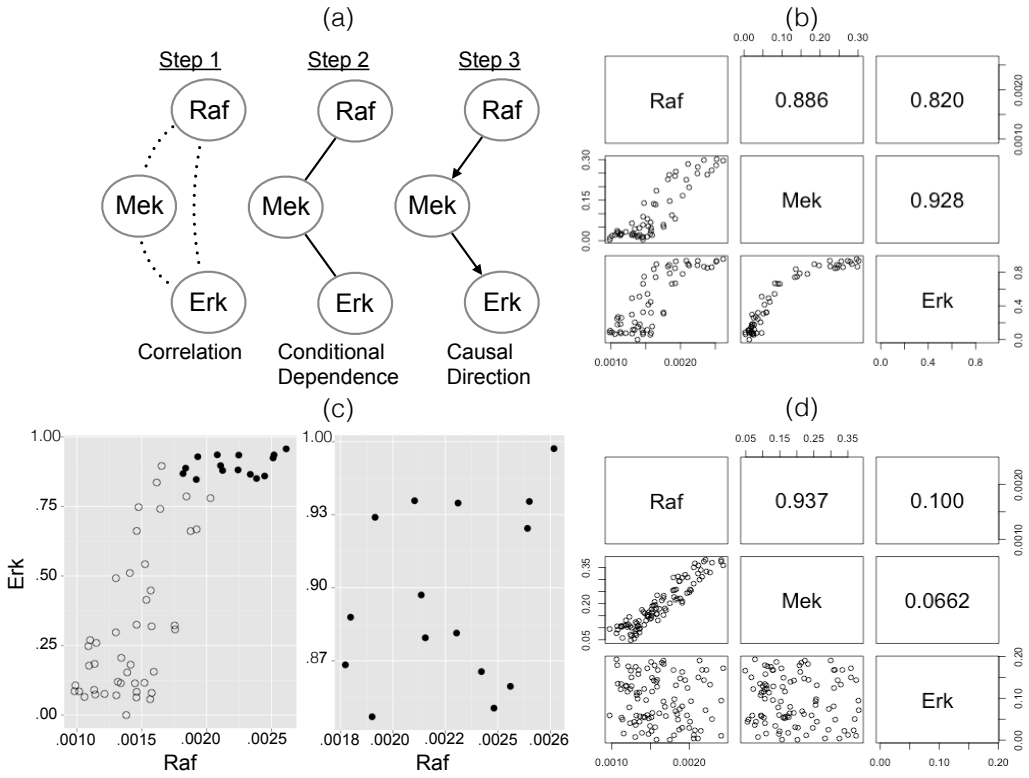

Fig. 2.2. (a) Overview of the 3 steps of causal inference, illustrated for the MAPK signaling cascade. (b), (c) and (d) feature an experiment simulated from the Huang-Ferrell computational model of the phosphorylation cascade. (b) Pairwise plots of concentration values of phosphorylated (doubly phosphorylated for Mek and Erk) forms of each protein, and observed Pearson correlations. The Raf – Erk correlation is high, despite the fact that Raf does not directly regulate Erk. (c) Concentrations of Raf versus Erk, where samples corresponding to high Mek (here, set to the top quartile) are highlighted with filled circles. The right panel of C shows the subset of samples with high Mek (i.e. conditional on Mek being high). In these samples the association between Raf and Erk disappears, and we infer that Raf and Erk are *conditionally independent* given Mek. (d) After Mek is inhibited, the observed association between Raf and Mek remains, while the association between Mek and Erk disappears. This reveals the causal flow from Raf to Mek. Figure republished with permission from original source [46].

proteins is comparable to the number of biological replicates, and step 3 requires systematic interventions (e.g. with protein inhibitors).

Figure 2.2 (b) illustrates Step 1 of the causal inference, and shows 2-way plots of the protein concentrations across the biological samples, and Pearson correlations to quantify the extent of the associations. The correlation values are high, and would meet most reasonable cut-off thresholds for constructing the correlation network in the left part of panel (a). The Raf–Mek and the Mek–Erk correlation edges match the Raf→Mek, Mek→Erk known causal edges. What about the non-causal Raf–Erk edge? Despite the high Raf–Erk correlation, there is no direct causal mechanism between them (aside from the one via Mek, which is already accounted for via the Raf→Mek and Mek→Erk edges). In causal inference, our goal is to eliminate this "nuisance" edge. How is this done?

To describe Step 2 of causal inference, we introduce some terminology. If concentrations of two proteins vary between the biological samples in a coordinated manner, such that knowing the concentration of one protein provides information on the concentration of the other, the two proteins are called *dependent*. Otherwise they are called *independent*. *Conditional independence* is a special case that is important to causal inference. Two *dependent* proteins are *conditionally independent* if, after knowing (in probability language, *conditioning on*) the concentration of third-party biomolecules, the two proteins become independent. In other words, after considering the information from the third party, the behavior of one protein provides no additional information on the behavior of the other.

While statistical associations and correlations are properties of the observed data, dependence and conditional independence are properties of the underlying processes that generate the data. Step 2 relies on statistical inference [9, 10] to infer from the observed data pairs of proteins that are conditionally independent. The conditionally independent pairs are ignored, and the remaining pairs are kept as hypothesized causal relations. The sparsity of the inferred conditional independence graph

is desirable, as it reduces the many pairwise associations to a smaller number of hypothesized causal relations.

Let's see how this applies to the MAPK signaling cascade. Since Raf regulates the concentration of Erk by way of regulating Mek, Raf and Erk are dependent. However, if we know the concentration of Mek, then the concentration of Raf provides no additional information about the concentration of Erk. Therefore, even though Raf and Erk are dependent, they are also conditionally independent given Mek. Figure 2.2 (c) illustrates the process of statistical inference by comparing the concentrations of Raf and Erk. When we subset the measurements to only the samples with high Mek, we can no longer see the association between Raf and Erk. Formally, the algorithm tests the null hypothesis of conditional independence between Raf and Erk given the full range of values of Mek (and not just high values of Mek, as was shown for the purposes of illustration), and evaluates evidence against the null hypothesis [27]. In this example the test did not reject the null hypothesis, and resulted in removing the edge between Raf and Erk as in the middle graph of Figure 2.1 (a).

At Step 2 the direction of the regulation remains unknown. Inference of the direction of the chain of events requires the experimental design, which involves external interventions or stresses. Figure 2.2 (d) illustrates the results of Step 3, in the case where an intervention targeted Mek with an inhibitor. The intervention does not affect the concentration of Mek, however it blocks its ability to phosphorylate other proteins. After this intervention the Raf–Mek relationship is unchanged, while Erk drops to a low level. From this we can infer that Mek has causal influence on Erk. Since Raf was unaffected by the intervention, Mek does not have a causal influence on Raf, and therefore the direction of causal influence in this edge goes from Raf to Mek. This intervention is required to transform the undirected graph in panel (a) - Step 2 to the causal graph in panel (a) - Step 3.

In the general case, computational methods for causal inference follow the workflow in Figure 2.2 (a), while scaling it to characterize multiple inter-related pro-

teins. Step 1 creates a dense network of pairwise associations. Step 2 identifies cases of conditional independence, and removes edges between conditionally independent proteins to create a much sparser network. Finally, Step 3 uses the experimental design, specifically the information regarding the interventions, to evaluate these edges as evidence for potential causal events. See Koller and Friedman [5] for a detailed description of these methods and their theoretical underpinnings. Numerous implementations are available in statistical software, e.g. in the R package `bnlearn` [28]. Depending on the biological system and on the experimental setting, the strength of causal evidence may vary. For example, Sachs, Itani *et al.* 2013 [21] highlight conditions in phoshoproteomic experiments where it may be infeasible to disentangle causality using perturbations.

2.3 Simulations

2.3.1 Large-scale statistical inference of causal relationships: challenges of scaling up

High-throughput proteomic experiments quantify a relatively large number of proteins (typically hundreds or thousands) as compared to the number of biological replicates (typically tens or hundreds). The large number of analytes creates challenges to the causal inference workflow.

In Step 1, the challenge is in accurately detecting statistical associations between pairs of the observed protein concentrations. A large number of quantified proteins produces a large number of spurious associations, which arise without any biological justification as an artifact of random chance. They obscure the systematic associations such as between Raf, Mek and Erk. Similarly, the spurious associations undermine Step 2, which starts from the correlation graph in Step 1, and eliminates the edges between conditionally independent proteins. More spurious associations lead to more undue rejections of conditional independence, and therefore to more false hypotheses of causal events.

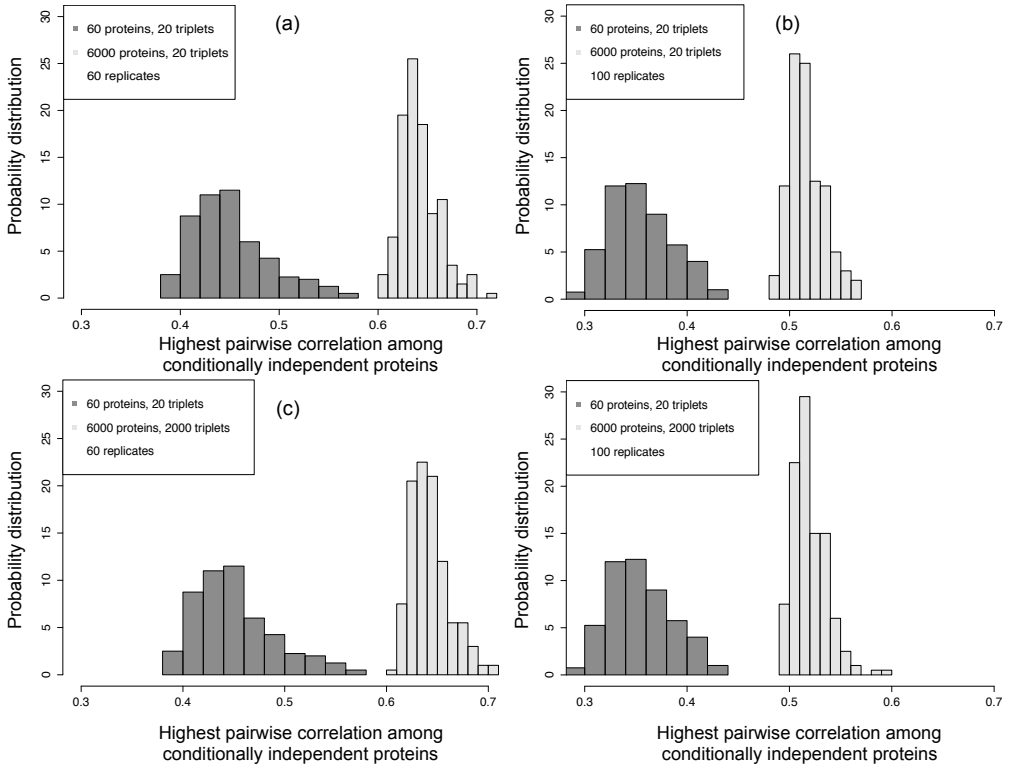

Fig. 2.3. Highest pairwise Pearson correlations among conditionally independent protein pairs, in 500 repetitions of the simulated experiments. (a) "Uninformative" increase in quantified proteins, smaller sample size. (b) "Uninformative" increase in quantified proteins, larger sample size. (c) "Informative" increase in quantified proteins, smaller sample size. (d) "Informative" increase in quantified proteins, larger sample size. Figure republished with permmission from original source [46]

To illustrate these challenges, we conducted a computer simulation inspired by Fan et. al [29], but translated to our context. The simulated experiments were purposely designed to be simple, in order to gain insight. We simulated two types of proteins. First, biologically "informative" proteins were represented by disjoint triplets such as Raf, Mek and Erk, where proteins within each triplet were statistically associated, but two proteins were conditionally independent given the third. Second, biologically "uninformative" proteins were simulated as independent (and therefore, conditionally independent) from every other protein. We studied the im-

pact of (1) the total number of quantified proteins, (2) the number of "informative" proteins, and (3) the number of biological replicates on the discovery of true pairwise associations in Step 1, and on the discovery of true conditional independence in Step 2. The details of the simulation are in Supplementary materials.

Figure 2.3 illustrates the challenges of Step 1 when working with high-throughput experiments. Figure 2.3 (a) presents a simulation that mimicked a targeted experiment, with 60 biological replicates and 60 proteins, all of which were "informative" and formed 20 triplets. It also presents a second simulation, which mimicked the worst-case scenario for a high-throughput experiment. It quantified 6,000 proteins, however all the newly quantified proteins were "uninformative". The simulations were repeated 500 times. Figure 2.3 (a) shows the probability distribution of the highest Pearson pairwise correlations among the conditionally independent proteins in each of the 500 repetitions. As can be seen, high-throughput experiments produce higher values of pairwise correlations, and therefore lead to more reported spurious associations.

Figure 2.3 (b) repeats the simulations in Figure 2.3 (a), while increasing the sample size to 100 biological replicates. It shows that increasing the sample size reduces the highest pairwise correlations among the conditionally independent proteins, and therefore helps minimize spurious associations.

Of course our expectation is that high-throughput experiments quantify more biologically informative proteins, and not just noise. We therefore repeating the simulations above, with the best-case scenario for a high-throughput experiment. It quantified 6,000 "informative" proteins that formed 2,000 triplets. Figures 2.3 (c) and (d) show the probability distribution of the highest Pearson pairwise correlations among the conditionally independent protein pairs in this scenario. As can be seen, the top and the bottom panels of Figure 2.3 are very similar. In other words, if the number of conditionally independent proteins in an experiment is relatively large, the number of true causal events has little impact on the extent of spurious associations.

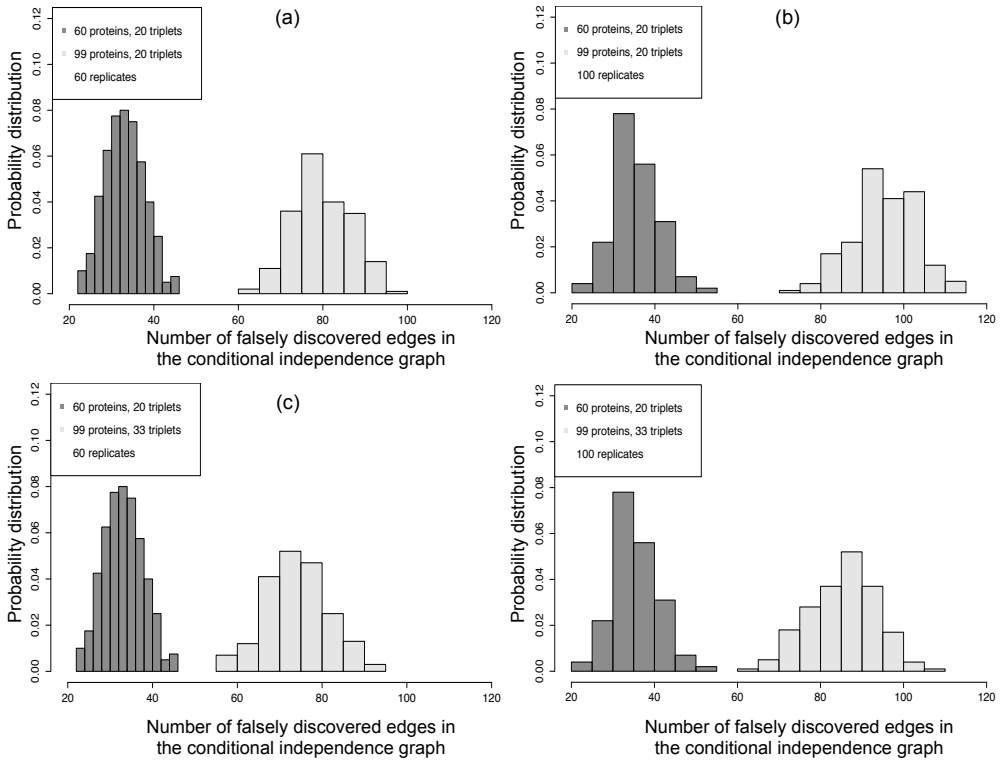

Fig. 2.4. Number of falsely discovered edges in a conditional independence graph, in 500 repetitions of the simulated experiments. (a) "Uninformative" increase in quantified proteins, smaller sample size. (b) "Uninformative" increase in quantified proteins, larger sample size. (c) "Informative" increase in quantified proteins, smaller sample size. (d) "Informative" increase in quantified proteins, larger sample size. Figure republished with permission from original source [46]

Figures 2.4 and 2.5 analyze the performance of Step 2 for the simulated experiments above. Since Step 2 is only applicable to experiments where the number of proteins is comparable to the number of biological replicates, we reduced the high-throughput experiment to 99 proteins. Despite the reduction, this step remains computationally intensive as it requires us to evaluate complex dependencies among 1,770 pairs of the 60 proteins, and 4,851 pairs among the 99 proteins. The simulation used the Grow-Shrink algorithm [30], which implements a greedy local constraint-based search [27] that iteratively revisits subsets of protein pairs and tests them for conditional independence.

Figure 2.4 (a) shows the probability distribution of falsely discovered edges in the conditional independence graph over the 500 repetitions. Since the high-throughput experiment starts with higher pairwise correlations in Step 1 and performs more tests, it produces more false edges, and therefore leads to more false hypotheses of causal events.

Figure 2.4 (b) shows the effect of increasing the sample size, and highlights the artifact of greedy local exploration of a complex high-dimensional space. A larger sample size increases the power of the individual tests for conditional independence, and allows the algorithm to explore larger subsets of protein pairs. More tests lead again to increasing the number of falsely discovered edges. Our additional simulations (not shown) illustrate that the number of falsely discovered edges stabilizes when the sample size is extremely large (tens of thousands). Figure 2.4 (c) and (d) show that including more "informative" proteins reduces the opportunity for false positives, but produces qualitatively similar results.

Figure 2.5 is the counterpart of Figure 2.4 that shows the main outcome of our interest, i.e. the ability to uncover edges that arise from true causal events. Figure 2.5 (a) shows that quantifying more "uninformative" proteins leads to fewer correct edges, and reduces the detection of correct edges as compared to the lower-throughput experiment. Figure 2.5 (b) shows that increasing the sample size improves the statistical power and leads to more correctly discovered edges, however

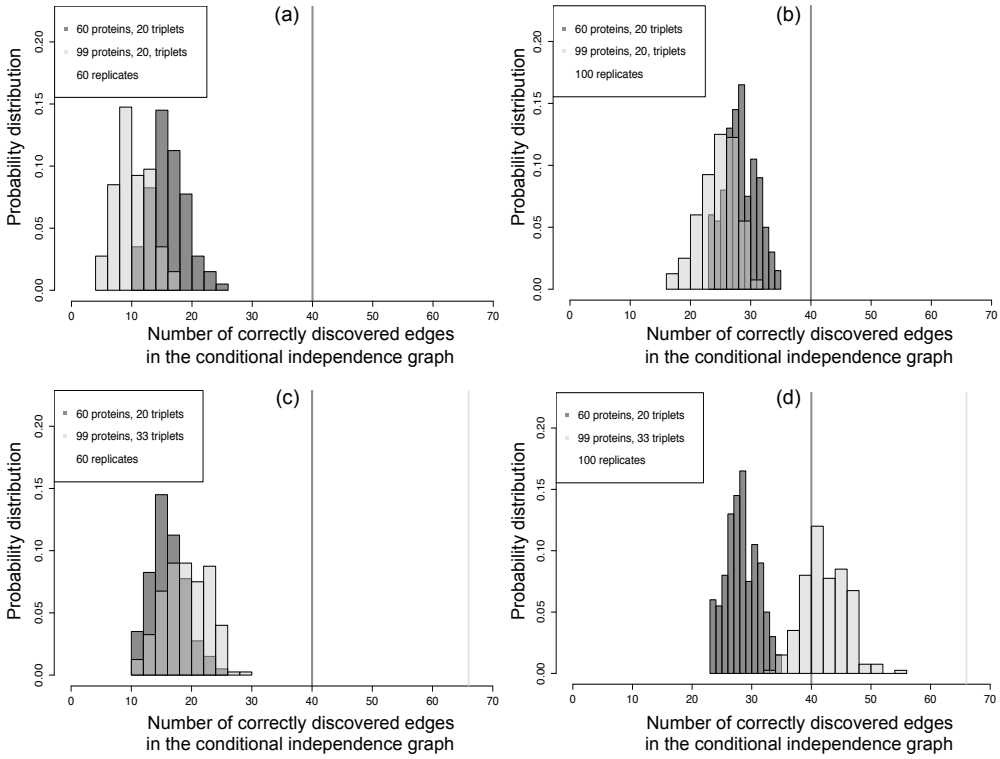

Fig. 2.5. Number of correctly discovered edges in a conditional independence graph, in 500 repetitions of the simulated experiments. Vertical lines indicate the true number of edges in each experiment. (a) "Uninformative" increase in quantified proteins, smaller sample size. (b) "Uninformative" increase in quantified proteins, larger sample size. (c) "Informative" increase in quantified proteins, smaller sample size. (d) "Informative" increase in quantified proteins, larger sample size. Figure republished with permission from original source [46]

the high-throughput experiment still performs worse. Figures 2.5 (c) and (d) show that the high-throughput experiment has the advantage in terms of discovering the correct edges when all the newly quantified proteins are "informative".

To verify the generality of these results, we repeated the analyses in Figures 2.3, 2.4 and 2.5 using the Incremental Association Markov Blanket algorithm [31, 32] (not shown), and obtained similar results.

Step 3 in the overall workflow is also challenged by an increased number of proteins. The first challenge is the large number of interventions required to infer the causal flow between protein pairs. Although only one intervention was sufficient in the Raf-Mek-Erk pathway, it is generally not enough. For example, if an experiment with k proteins can only perturb one protein at a time, $k-1$ interventions are required to infer causality. If an experiment can simultaneously perturb multiple proteins the number of interventions can be reduced [12], however success of these interventions depends on multiple complex factors [33, 34]. The second challenge is the presence of false putative causal relations inherited from Steps 1 and 2, which will adversely affect Step 3 regardless of the size of the dataset.

2.4 Discussion

2.5 Approaches for inferring causality from high-throughput experiments

The problems outlined above paint a grim picture for causal inference in large datasets. Fortunately, these can be overcome, and effective causal inference can be a reality for high-throughput experiments. We provide suggestions for the best practices below.

1. *Limit the number of analytes.* Reducing the number of proteins minimizes false hypotheses, and improves our ability to correctly hypothesize causal events. As the technologies improve the lists of quantified proteins grow larger, however only a subset of the measurements is both biologically relevant and technologically precise. The length of the list is not an important indication of the performance of the experiment. If the broader biological system is well understood, it is possible to design a targeted experiment that focuses on a specific network or pathway, and ask more specific questions of the data, such as the presence of a particular regulatory event. The more specific the question, the less data are needed to make solid causal conclusions.

2. *Profile more biological replicates.* Increasing the number of replicates is an effective strategy for both reducing spurious pairwise correlations, and improve our ability to correctly hypothesize causal events. Therefore, the high-throughput experiments should include more samples from distinct biological sources, i.e. from distinct biological individuals representing the underlying population. Most current technologies requires a trade-off between the number of quantified proteins and the number of replicates [35]. Recent mass spectrometric technologies such as Data Independent Acquisition or SWATH [36] have a high promise of expanding the limits of this trade-off. At the same time, single cell mass cytometry quantifies thousands of cells per biological sample and provide rich input to causal inference [37]. However these data should be used with caution, as multiple cells represent a single biological individual. Data from multiple individuals are required to make broader inference regarding the underlying population.

3. *Use prior knowledge.* Prior knowledge improves the search for conditional independence and helps to determine causality. The prior knowledge can be used in form of known canonical networks, extracted, e.g. from pathway databases such as KEGG. One example of such prior information is the MAPK pathway. The prior information reduces the search space of unknown associations that need to be considered, enables a more effective use of the data, and increases the confidence in newly discovered statistical associations. Saez-Rodriguez, Lauffenburger, and Sorger [38], as well as Terfve and Saez-Rodriguez [39] use prior network knowledge to build logic models that reflect causal relationships between signaling proteins from protein concentrations. Another example of prior knowledge is contextual information, such as spatial or temporal annotations of the quantitative measurements in the cell. The contextual information can be extracted from the literature or from other complementary (and potentially noisy) datasets. The causal inference algorithms can be ex-

tended to weigh evidence of conditional independence, depending on whether the proteins share the same spatial or temporal context.

4. *Select targeted interventions wisely.* Targeted interventions perturb individual components of the biological system. An example is siRNA knockdowns, as well as small molecule inhibitors, which block the causal influence of a specific protein on its downstream components. Although effective, such targeted interventions are limited in number. Therefore, a strategic experimental design would use prior information, prioritize the interventions and the targets, and apply them to parts of the biological system that have most potential for new discovery of regulatory events. For example, a graph with undirected edges can be inspected, to reveal which nodes have potential to reveal the most causality if perturbed. Such targeted perturbations can be applied iteratively, after an initial statistical analysis revealed areas of the network where causal inference would benefit from extra measurements and data.

5. *Consider broad-scale interventions.* Broad-scale interventions sacrifice specificity of targeted interventions to simultaneously perturb many proteins in a biological system. One example of broad-scale interventions is varying experimental conditions, in order to activate multiple pathways. Signals from endocrine, paracrine, and autocrine ligands elicit various signaling responses in hepatocytes, thus interventions that cover this range of signals gives the best picture of the broader causal network of hepatocyte signaling [40]. Similarly, interventions that go beyond receptor-level and perturb multiple components of the system bring cascading causal direct orientation deeper into the network. Although they do not provide specific information about the downstream effects of stimulation, broad-scale interventions can provide more causal insight. Therefore, the advantage of this approach is that it may enable elucidation of causality across the entire system.

This list suggests impactful approaches that can drastically improve causal inference from high-throughput experiments. They provide various constrains on the inference task, thereby improving the accuracy of the conclusions. For example, the task of assessing which of all the possible KEGG pathways is present in the dataset is far less error-prone than the task of assessing which of all possible combinations of the quantified proteins might form a biological pathway. These approaches are most powerful when used in combination, and in fact the lines between them are somewhat arbitrary and frequently blurred. For example, using items 1 and 2 in concert can be thought of as reducing the breadth and increasing the depth of the investigation. Items 4 and 5 call for use of interventions, but this task itself is complicated by measuring many proteins. Item 3, prior biological knowledge, can be used to prioritize what to target with that limited set of interventions. Causal inference becomes possible when combining these tools within a sound experimental design.

3. A BAYESIAN ACTIVE LEARNING EXPERIMENTAL DESIGN TO INFER SIGNALING NETWORKS

3.1 Introduction

Signaling networks describe chains of protein interactions that determine how cells process signals from their environment. The deregulation of signaling networks occurs under many conditions, e.g. in diseases such as cancer [41], gene knockouts, or introduction of a drug. The patterns of such deregulation can be inferred from quantitative proteomic experiments conducted under the conditions of interest, using causal inference and Bayesian networks [8, 11].

In these investigations, signaling is induced with a stimulus perturbation, and a measurement technology acquires information on the activity of signaling proteins [2]. Bulk experiments quantify aggregate signaling activity across a sample. In contrast, single cell technologies provide cell-level resolution of signaling activity. For example, in flow cytometry cells are chemically fixed, and intracellular signaling proteins are tagged with fluorescently-labeled antibodies. The cytometer then records the antibodies' fluorescence in individual cells, each reflecting the relative abundance of signaling proteins in different states of enzymatic activity [42]. Similarly, in mass cytometry (CyTOF) experiments, intracellular signaling proteins are tagged with heavy-metal isotopes, and the mass spectrometer records the mass-to-charge ratio of the charged isotope tags [43].

Causal Bayesian networks represent signaling proteins as nodes, and regulatory relationships as directed edges. It interprets a network as a topological map of the underlying signaling network. By comparing the structure inferred under a condition to *canonical pathways* in sources such as KEGG and Reactome, we can learn the patterns of network deregulation. Data repositories, such as Cytobank [44], pro-

vide *historic data*, which can be incorporated in the analysis and interpretation of experimental results [8,45]. This prior information is especially important for higher throughput experiments because it helps eliminate spurious correlations and false discoveries of relationships between proteins [46].

To distinguish causal relationships from statistical associations, causal network inference requires *targeted interventions* on some proteins [4, 46], e.g, using small-molecule inhibitors that block a protein's enzymatic activity. An insufficient set of interventions results in only a partially causally oriented network [6]. At the same time, increasing the number of interventions increases the complexity of the experiment and the cost. This cost is wasted when targeted interventions redundantly orient the same edges.

In this paper, we propose a strategy for optimal design of bulk or single-cell proteomic experiments aiming at causal inference. The design prioritizes targeted interventions, and provides a criterion to stop adding interventions. It combines prior knowledge in the form of canonical pathways imported from sources such as KEGG [20] with historic data. The strategy outputs a sequence of interventions that we call a "batch", i.e. a minimal subset of candidate interventions that contributes maximal causal information given the available data. We then describe an active learning framework, that iterates between selecting interventions and acquiring data to obtain a fully inferred causal network. To the best of our knowledge, this is the first active learning approach to experimental design for inference of signaling networks.

3.2 Background

3.2.1 Directed graphs as causal models of signaling

A causal Bayesian network denotes a set of p signaling proteins with p nodes $V = \{v_1, ..., v_p\}$. The nodes are variables representing levels of signaling activity of the proteins. For example, v_1 can take discrete signaling states "active" or "inactive",

or continuous values quantifying the abundance of a protein form. The model expresses causal relations between nodes with a directed acyclic graph structure (DAG) G. The edge direction in the DAG represents the causal effect of a change in the signaling state of a parent node on the state of the child. The DAG is best interpreted as a snapshot of a dynamic system [3]. This interpretation is strongest when the signaling response has reached some quasi-steady-state.

Each node in the DAG has a conditional probability distribution given its parents. It is a probabilistic representation of the regulatory influences of the parents on the child [10]. A key advantage of the probabilistic interpretation is that it encodes *conditional independence*, i.e. the probability that the state of a protein is independent of the state of all its upstream proteins, if we know (i.e. condition on) the states of its direct parents. This allows us to ignore the correlation between a protein and proteins more than one step upstream.

From the statistical perspective, the goal of causal inference is to infer the DAG structure representing the signaling network, using associations between proteins as input. However, statistical associations are not sufficient to orient the edges in the DAG [46]. We illustrate this with the simple 3-protein canonical MAPK signaling pathway Raf $\rightarrow$ Mek $\rightarrow$ Erk. Imagine that the structure of the pathway is unknown, and needs to be inferred. A causal inference algorithm would (1) detect pairwise statistical correlations between abundances of each pair of the three proteins, pointing to the three candidate edges, (2) test Raf and Erk for conditional independence, given the state of Mek, and (3) in presence of conditional independence, eliminate the edge between Raf and Erk. After that, additional interventions are required to orient the edges between Raf and Mek, and between Mek and Erk. The left box in Figure 4.4 illustrates that, in absence of interventions, the ground truth is statistically indistinguishable from the other two causally incorrect DAGs. A set of statistically indistinguishable DAGs form a *Markov equivalence class* P, comprised of DAGs with same edges but varying orientations, which have equal statistical likelihood for the dataset [5, 6]. A Markov equivalence class is repre-

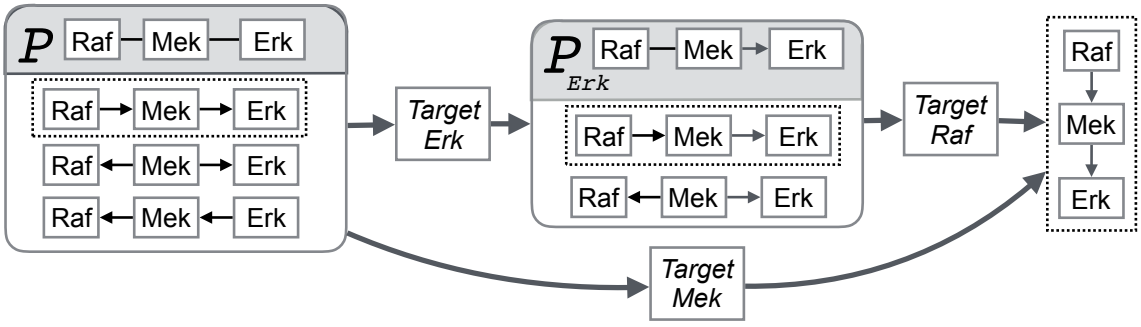

Fig. 3.1. Illustration of DAG equivalence classes. The DAG Raf $\rightarrow$ Mek $\rightarrow$ Erk is the "ground truth" canonical MAPK signaling pathway, which we seek to learn by causal inference. The left box shows the equivalence class P, represented by the PDAG Raf $-$ Mek $-$ Erk. The PDAG contains three DAGS, all statistically indistinguishable in absence of interventions. The cardinality of P is 3. The middle box shows the PDAG P_{Erk}, obtained after an intervention or Erk. P_{Erk} is a subclass of P that has eliminated Raf $\leftarrow$ Mek $\leftarrow$ Erk. The cardinality of P_{Erk} is 2. The right box shows the single ground truth DAG obtained after an additional intervention on Raf. An alternative single intervention on Mek simultaneously compels the direction of both edges, and is more effective at discovering the ground truth.

sented with a partially directed acyclic graph (PDAG). The directed edges in a PDAG have the same orientation in all the DAGs in P. The undirected edges in the PDAG have varying directions in the DAGs, and therefore represent edges with uncertain causality. Figure 4.4 illustrates PDAGs and DAGs in the MAPK pathway example. The cardinality of a PDAG is defined as the number of its DAGs.

Targeted interventions proceed by fixing the state of a node, such that it does not vary with that of its parents [4, 5], [6]. Fixing the node introduces an additional constraint, and eliminates members of the equivalence class that fail to satisfy this constraint. For example, in Figure 4.4 a small inhibitor fixes Erk's enzymatic activity state to "off", such that the activity of Erk becomes independent of the state of Mek. The intervention fails to regulate the activity of Mek, and therefore eliminates the DAG with an edge Erk $\rightarrow$ Mek.

The reduced equivalence class is formally defined as a *transition-sequence Markov equivalence class*, i.e. the equivalence class after a sequence of "transitions" (interventions) [17]. Each additional intervention orients more edges in a PDAG, and a sufficiently large set of interventions compels all the edges. In Figure 4.4, interventions on Erk and Raf eliminated all but the ground truth from the equivalence class.

As the example illustrates, a batch of interventions targeting Erk and Raf would reveal the ground truth DAG. However, so would a batch containing a single intervention on Mek. The goal of this work is to identify batches of interventions, which reveal the most causality while minimizing the number of interventions they contain, and by extension the experimental complexity and cost.

3.2.2 Bayesian inference of causal networks

This work focuses on a Bayesian approach to learning causal PDAG representations of signaling, where experimentalists (1) start with background knowledge about the signaling network, such as likely pathways or motifs, (2) use the background knowl-

edge to construct a *prior* distribution on graph structures, (3) collect experimental measurements of the signaling states of proteins, and (4) estimate a *posterior* distribution on structures based on the prior and the experimental measurements. The Bayesian approach is advantageous, as it reveals the "rewiring" of signaling between conditions by examining differences between the prior and the posterior distributions.

More formally, the approach uses the background knowledge δ to construct a prior probability distribution of possible DAG structures $\pi(G|\delta)$ in the space of possible structures $\mathbb{G}$. The experimentalist collects a data set D of measurements on the proteins $V \in G$, acquired under a batch of targeted interventions S. A statistical likelihood function $p(D|S, G)$ quantifies the likelihood that the observations were generated from a graph G. Finally, inference of the DAG structure relies on a posterior probability distribution $\pi(G|D, S, \delta)$

$$\pi(G|D, S, \delta) \propto p(D|S, G)\pi(G|\delta) \tag{3.1}$$

In many biological applications, inferring a single DAG with high posterior probability is not of the main interest. Instead, we are interested in local features of the graph, such as the presence of particular edges or network motifs. This is expressed through a function f on a graph that quantifies the feature of interest, and through its posterior expectation $E\{f|D, S, \delta\}$ across all graphs.

$$E\{f|D, S, \delta\} = \sum_{\mathbb{G}} f(G)\pi(G|D, S, \delta) \tag{3.2}$$

For example, if f is an indicator of the presence of an edge in the network, then $E\{f|D, S, \delta\}$ is the posterior probability of the presence of that edge. In this work, our feature of interest quantifies the causal insight provided by an intervention.

Bayesian inference of DAG structures relies on a *Bayesian scoring function* $Score(G, D, S, \delta)$ that takes as arguments a DAG, a dataset quantifying protein activity, a set of interventions, and the structured prior knowledge. It returns values proportional to the posterior distribution $\pi(G|D, S, \delta)$. Algorithms searching for DAGs with high

$Score(G, D, S, \delta)$ must explore the combinatorially large search space of possible DAGs [30, 47, 48]. For computational tractability, some algorithms approximate the posterior probabilities $\pi(G|D, S, \delta)$ (see [5, 49] for background, and [50, 51] for an example). Since the full search space over all possible DAGs $\mathbb{G}$ is combinatorially large, the common practice is to sample a set of DAGs from their posterior distribution through a random process such as bootstrap ([52, 53]) or MCMC [54], and keep a sample of high scoring networks. $E\{f|D, S, \delta\}$ is then approximated as

$$E\{f|D, S, \delta\} \approx \sum_{G \in \Omega} f(G) \frac{Score(G, D, S, \delta)}{\sum_{G; G \in \Omega} Score(G, D, S, \delta)} \tag{3.3}$$

where Ω denotes a sample of high scoring DAGs.

3.2.3 PDAG representation of uncertainty in causal effects

Experiments with incomplete sets of interventions lack information to fully infer the causal orientation of the edges in a DAG. In order to characterize this uncertainty, algorithms take a single DAG and determine the space of DAGs having the same conditional independence structure, topology and likelihood $p(D|S, G)$, but different edge orientations as the input DAG [17, 55]. These DAGs form a *Markov equivalence class*. The algorithms return a PDAG, which represents the class by preserving the shared edge structure, while presenting edges with conflicting orientations as undirected.

In Bayesian inference of causal networks, we are interested in Markov equivalent graphs with not only the same likelihood, but also the same posterior probability and the same score $Score(G, D, S, \delta)$ [55]. As seen from Equation 3.1, Markov equivalent graphs have the same posterior only if they have the same prior probability $\pi(G|\delta)$. This last condition does not hold when the prior probabilities encode available causal information, such that for some edges orientation in one direction is more probable than the other. A contribution of this manuscript is an implementation of a DAG-to-PDAG conversion algorithm that accounts for informative prior

probabilities of causal direction of edges, and outputs a PDAG representing a class of DAGs that are Markov equivalent and have the same posterior and score.

3.2.4 Active learning for the optimal design of causal network inference experiments

In the context of causal inference from experiments, active learning is the task of including targeted interventions in the design, to optimize the inference of edge orientation. For example, in Figure 4.4 an intervention on Mek compels both edges in the graph, and is thus more valuable than an intervention on Erk, which only orients one edge.

Previous work used active learning to distinguish members of statistically equivalent graphs [12–14], [16, 17]. However, these approaches work with either a single PDAG, or a selection of highly-likely PDAGs. The approach in this manuscript differs by working with the entire probability distribution of PDAGs, characterizing each PDAG by its posterior probability of containing the causal truth.

Alternative Bayesian approaches to active learning of causal networks also exist [15, 18]. However, they require the experimentalists to represent their background knowledge in terms of topological orderings, i.e. an ordering of nodes such that the "from" node for every edge occurs earlier in the ordering of the "to" node. In contrast, this manuscript represents background knowledge in terms of probabilities of edge presence and orientation. This more intuitive approach simplifies the experimentalists' work with pathways.

Finally, the proposed approach is similar in spirit to other methods in the bioinformatics literature that use historic data to inform experimental design. E.g., Rossel and Muller [56] used a sequential Bayesian method to plan sample size. Guan et al. [57] used available data to find optimal orderings of high-throughput experiments. King et al. [58] constructed a "robot scientist" that applied an active-learning strat-

egy to functional genomics. To our knowledge, this manuscript is the first to apply active learning to inferring the structure of cell signaling pathways.

3.3 Methods

3.3.1 Prior knowledge for causal graph structure learning

Quantifying causal knowledge with edge probabilities. We propose to use probabilities of edge presence and edge orientation as a means of modeling signaling events. A set of signaling proteins, represented by nodes in V, has up to $\frac{|V|(|V|-1)}{2}$ possible edges. *Presence probability* of an edge between nodes $\{u, v\}$, denoted $P(u-v)$ or $\overline{\pi}_{uv}$ as a shorthand, quantifies the confidence that the edge is present in G. *Orientation probability* for the edge from u to v, denoted $P(u \rightarrow v|u-v)$ or $\overrightarrow{\pi}_{uv}$ as a shorthand, is the conditional probability of this orientation, *given* that the edge is present. Since only two orientations are possible, $P(u \rightarrow v|u-v) = 1 - P(u \leftarrow v|u-v)$. The goal of targeted interventions is to resolve the orientations of the edges, i.e. coerce orientation probabilities towards 0 or 1.

Let δ be a set of edge probabilities $\delta_{u,v}$, where

$$\delta_{u,v} \stackrel{\text{def}}{=} \{\overline{\pi}_{uv}, \overrightarrow{\pi}_{uv}\} \tag{3.4}$$

In Bayesian setting, we wish to use the edge probabilities to quantify prior causal knowledge. Let $\overline{I}_{uv}(G)$ be an indicator function for the presence an an edge between u and v in G, and $\overrightarrow{I}_{uv}(G)$ be an indicator function for the orientation of the edge from u to v. We map edge probabilities δ to a probability distribution on DAG structures using using an *edge-wise prior*

$$\pi(G|\delta) = c \prod_{uv \in G; u \neq v} \left(1 - \overline{\pi}_{uv}\right)^{1-\overline{I}_{uv}(G)} (\overline{\pi}_{uv} \overrightarrow{\pi}_{uv})^{\overline{I}_{uv}(G)\overrightarrow{I}_{uv}(G)} \tag{3.5}$$

where c is a normalizing function on the space of graphs that corrects for the acyclicity constraint [59].

When nothing is known about the presence or orientation of the edges, we specify the *uninformative* edge probabilities [49], where

$$\overline{\pi}_{uv} \approx \frac{1}{2} + \frac{1}{2(|V|-1)}, \quad \text{and} \quad \overrightarrow{\pi}_{uv} = \frac{1}{2} \tag{3.6}$$

The intuition behind Equation 3.6 is that two nodes are more likely to be linked in a small network than in a large network. Therefore, the uninformative presence probability approaches .5 as network size increases. The uninformative orientation probability is .5 for either direction. These uninformative edge-wise priors correspond to the marginal probabilities of an edge in the case when there is a uniform probability distribution on the space of graphs [49].

Upon conducting an experiment with interventions S and collecting a dataset D, the next step is to *update* the DAG probability distribution with condition-specific information in the data using Bayes rule

$$\pi(G|D,S,\delta) \propto p(D|S,G)\pi(G|\delta) \tag{3.7}$$

Note that the Bayes rule is agnostic of the process that selected the interventions in S. S can be selected by any approach, such as applying the available inhibitors, or using the proposed active learning approach below.

The updated probability distribution $\pi(G|D,S,\delta)$ maps back to edge probabilities using the approach in Equation 3.2. Let $f(.)$ in Equation 3.2 be the indicator functions $\overline{I}_{uv}(G)$ and $\overrightarrow{I}_{uv}(G)$. Then, the updated presence and orientation probabilities of an edge after observing data D is defined as

$$\begin{aligned} \overline{\pi}_{uv|D,S} &= \sum_{\mathbb{G}} \overline{I}_{uv}(G)\pi(G|D,S,\delta) \\ \overrightarrow{\pi}_{uv|D,S} &= \frac{1}{\overline{\pi}_{uv}} \sum_{\mathbb{G}} \overrightarrow{I}_{uv}(G)\pi(G|D,S,\delta) \end{aligned} \tag{3.8}$$

In other words, these are average frequencies of edge presence and orientation over all the DAGs, weighted by the posterior probabilities of the DAGs.

Incorporating pathway knowledge and historic data. When prior information is available, we propose to construct informative edge probabilities $\delta_{u,v}$. In the

simplest case, a directed edge between u and v in the canonical pathway is viewed as a hypothesis that an edge linking these nodes is also present under the condition of interest, and is oriented from u to v. Denoting the set of edges in the canonical pathway as $\mathbb{K}$, the background knowledge $\delta_{u,v}$ is defined as in Equation 3.4 where

$$\overline{\pi}_{uv} = \begin{cases} \sim 1 \text{ if } u - v \in \mathbb{K} \\ \sim 0 \text{ otherwise} \end{cases} \tag{3.9}$$

$$\overrightarrow{\pi}_{uv} = \begin{cases} \sim 1 \text{ if } u - v \text{ is oriented } u \to v \in \mathbb{K} | u - v \in \mathbb{K} \\ \sim 0 \text{ if } u - v \text{ is oriented } u \leftarrow v \in \mathbb{K} | u - v \in \mathbb{K} \end{cases}$$

The notation ~ 1 (probability near 1) and ~ 0 (probability near 0) emphasizes that the Bayesian approach avoids the boundary probabilities of 0 and 1.

In some cases, experimentalists may wish to use alternative specifications. For example, in absence of canonical pathway information it may be inappropriate to assign $\overline{\pi}_{uv} \sim 0$ to each edge, and subjective edge probabilities may be a better choice. For edges where no assessments can be made, we use the uninformative edge probabilities in Equation 3.6.

In addition to incorporating prior knowledge from canonical pathways, we also seek to make use of information in historic datasets. We define historic data D_0 as previous experiments, which quantified the activity of the proteins in the same network, under the same signaling conditions as the pending causal inference experiment, but lacking targeted interventions. We update the canonical knowledge $\pi(G|\delta)$ with the condition-specific information in the historic data

$$\pi(G|D_0, \delta) \propto p(D_0|G)\pi(G|\delta) \tag{3.10}$$

Here $p(D_0|G)$ is the likelihood that the historic data came from the graph G, and $\pi(G|D_0, \delta)$ is the updated distribution on graph structures, which now captures the full state of our prior information before the interventions.

Sampling DAGS from a DAG distribution Similarly to Equations 3.2 and 3.3, the proposed Bayesian inference on causal networks relies on sampling a set of

DAGs Ω from a distribution $\pi(G|D, S, \delta)$. Our implementation uses Bayesian bootstrap sampling from a distribution of DAGs [52] and [53] with random graph starts [60] and greedy search, and the posterior distributions are derived using Bayesian Dirichlet approximation [51]. When the signal-to-noise ratio in the historic data is high and/or the edge-wise prior is informative, the sampling concentrates on a smaller set of most probable DAGs. When the signal-to-noise ratio is low, weight is distributed more evenly among graphs, and the sampling must cover a larger number of graphs with similar $\pi(G|D, S, \delta)$.

Representing uncertainty in causal effects with PDAGs. We incorporate the informative edge orientation probabilities in δ to express the uncertainty in edge orientation using PDAGs. We view the conversion of a DAG to a PDAG as a the function f in Equations 3.2 and 3.3. Applying these equations requires that DAG members of the same equivalence class have the same posterior probability, otherwise different instances of the same PDAG would have different probabilities, i.e. the same f would have different probabilities in Equation 3.2 and scores in 3.3. Current conversion algorithms are not compatible with edge-wise prior probabilities.

Algorithm 1 is a DAG-to-PDAG conversion algorithm that incorporates informative edge-wise prior probabilities. It starts with a DAG from Ω. Next, every directed edge is converted to an undirected edge if it meets three conditions. The first two conditions are the same as in the prior literature [17, 55]. First (lines 4 and 10 in Algorithm 1), reversing edge direction should not change the number of immoral v-structures (i.e., 3-node motifs with one child and two parents, with no edge between parents). Second (line 6 of Algorithm 1), the child node of the edge should not be targeted by an intervention in S. In this manuscript we introduce a third condition (lines 8 in Algorithm 1), stating that the edge should have an uninformative prior orientation probability of .5. These conditions create an equivalence class P and its PDAG, where all the members have the same value for $\pi(G|D_0, \delta)$. If the algorithm

is applied to two Markov equivalent DAGs with different causal information in their informative edge-wise priors, two different PDAGs are returned.

Algorithm 1 *DAG to PDAG Algorithm*

Inputs: A DAG G, an (optional) set of selected intervention targets S, a set of edge probabilities δ.

```
1:  procedure PDAG(G, S, δ)
2:      for edge e in G do
3:          if e is in an immoral v-structure then
4:              Fix direction of e
5:          if e's child is targeted by S then
6:              Fix direction of e
7:          if e's orientation probability in δ ≠ .5 then
8:              Fix direction of e
9:      for edge e in G do
10:         if e is not fixed then
11:             if reversing e's direction
                        will not add a new v-structure
                         or introduce a cycle then
12:                 Make e undirected
13:     P ← G
14:     return (P)
```

3.3.2 Bayesian active learning with causal information gain

Overview. The strategy for selecting optimal interventions is overviewed in Figure 3.2. The active learning algorithm takes as input a sample of DAGs from a DAG distribution, and a set of candidate interventions. The algorithm sequentially evaluates the expected causal information gain of the interventions, and outputs

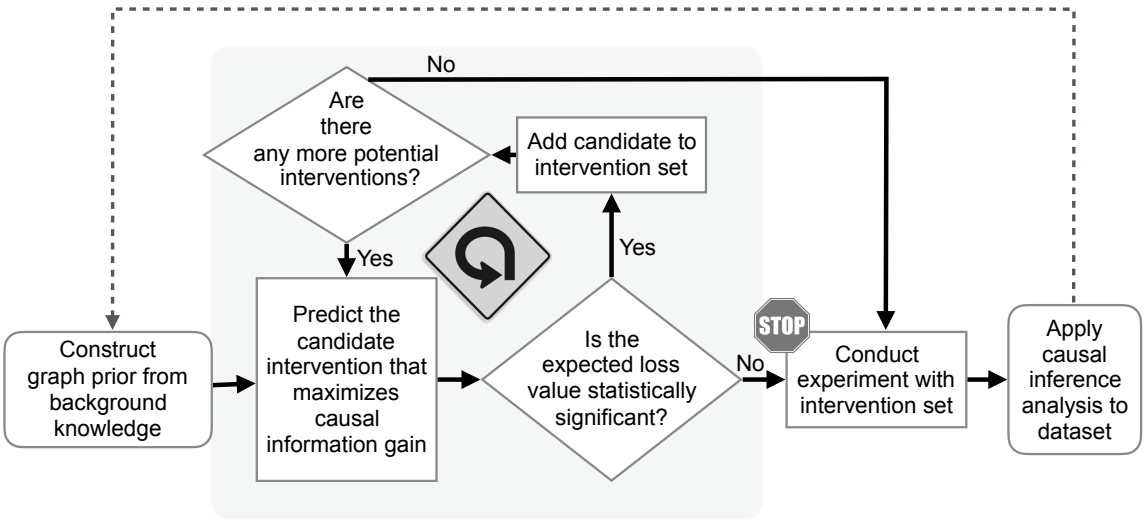

Fig. 3.2. Overview of the proposed method. A probability distribution of possible graph structures is constructed from canonical pathways and historic data. Interventions are iteratively added to the design until the expect causal information gain we can expect from an additional intervention becomes small. We then stop adding interventions to the design and use the newly acquired data to infer the causal network.

a minimally-sized batch of interventions that maximizes the expected information gain. We detail the components of the algorithm below.

Defining the causal information gain of an intervention. Suppose that the true causal DAG G were known. Let $S \subseteq V$ denote a batch of candidate interventions. As discussed in Section 3.2.3, a PDAG is derived directly from a DAG's topology and a set of interventions, and can be determined before collecting data. Therefore, we could devise an algorithm $H(G, S)$ that first determines a PDAG, then counts the number of oriented edges in the PDAG, i.e. the number of oriented edges in G that could be inferred from data with interventions S.
Suppose that we consider an additional intervention on node v, which leads to $H(G, \{S, v\})$ edges correctly oriented edges. We define the causal *information gain* $IG(G, S, v)$ of the intervention on v as the increase in correctly oriented edges, i.e. $IG(G, S, v) = H(G, \{S, v\}) - H(G, S)$. $IG(G, S, v)$ is non-negative, and can be zero if v fails to

orient any edges beyond those oriented by S. Note that an equivalent definition of information gain is the reduction in the number or unoriented edges. This definition parallels the information theory notation of entropy, where the information gain is viewed as entropy reduction.

Selecting interventions that maximize the expected information gain Of course in practice the true causal DAG G is unknown. Therefore, similarly to Equation 3.2 we consider the *expected information gain*, which averages the information gain over all possible graphs, weighted by their prior distribution $\pi(G|D_0, \delta)$

$$EIG_{S,v} = \sum_{\mathbb{G}} IG(G, S, v)\pi(G|D_0, \delta) \tag{3.11}$$

Moreover, similarly to Equation 3.3, we approximate the expected information gain as

$$EIG_{S,v} \approx \sum_{G \in \Omega} IG(G, S, v) \frac{Score(G, D_0, S, v, \delta)}{\sum_{G;G\in\Omega} Score(G, D_0, S, v, \delta)} \tag{3.12}$$

where Ω denotes a sample of of high scoring DAGs, and $Score$ is a Bayesian scoring function returning a value proportional to $\pi(G|D_0, \delta)$. We then select the candidate v that maximizes the approximated expected information gain. Algorithm 2 details these steps. Note that in Bayesian decision theory, $-IG$ is a loss function, and we select the candidate v that minimizes the expected loss [61].

Starting point, iterations and stopping criteria The proposed active learning strategy is summarized in Algorithm 3. It starts with the empty set of selected interventions $S = \emptyset$, a set of candidate interventions U, and a set of DAGs Ω. For each intervention $v \in U$ and each DAG G in Ω, the EIG$(\Omega, Score, S, v)$ algorithm calculates P_S, $P_{S,v}$ and $IG(G, S, v)$, and returns the expected information gain. The candidate with the maximum expected information gain is added to the batch S. In the next iteration, the expected information gain for the remaining candidates is evaluated, while accounting for the effect of the interventions that are already in the batch.

Algorithm 2 *Expected Information Gain*

Inputs: A set of DAGs Ω, Bayesian scoring function $Score$, a set of pre-selected intervention targets S, a candidate for next intervention v, and a set of edge probabilities δ.

1: **procedure** EIG(Ω, $Score$, S, v, δ)
2: Initialize array IG of size $|\Omega|$
3: **for** i in 1:$|\Omega|$ **do**
4: $P_S \leftarrow \text{PDAG}(G_i, S, \delta)$
5: $P_{S,v} \leftarrow \text{PDAG}(G_i, \{S, v\}, \delta)$
6: $H_S \leftarrow$ num. of directed edges in P_S
7: $H_{S,v} \leftarrow$ num. of directed edges in $P_{S,v}$
8: $IG_i \leftarrow H_{S,v} - H_S$
9: **return** WeightedMean(IG, $Score$)

Algorithm 3 *Active Learning*

Inputs: A set of DAGs Ω, Bayesian scoring function $Score$, a set U of candidates for next intervention, and a set of edge probabilities δ.

Parameter: α, stopping criterion for the information gain.

```
1: procedure ActiveLearning(Ω, Score, U, α, δ)
2:     S ← null
3:     while length(U) > 0 do
4:         TopCandidate ← null
5:         MaxEIG ← 0
6:         for node v ∈ U do
7:             EIG_{S,v} ← EIG(Ω, Score, S, v, δ)
8:             if EIG_{S,v} > MaxEIG then
9:                 TopCandidate ← v
10:                MaxEIG ← EIG_{S,v}
11:        if ¬ Stop(TopCandidate, Ω, α) then
12:            S ← cat(S, TopCandidate)
13:            U ← U[-TopCandidate]
14:        else
15:            return (S)
```

After a certain point, additional interventions to the batch become counterproductive. For example, in Figure 4.4 an intervention on Mek would orient the $Mek - Erk$ edge. Including an additional intervention on Erk would provide no additional information. In the case of Figure 4.4 the true graph is known, and we stop adding interventions when the information gain is 0. Since in real-life situations the structure of the true DAG is unknown, we stop adding interventions when the probability that at least some information gain occurs is below a parameter α. Similarly to Equations 3.3 and 3.12, the probability that at least some information gain occurs is

$$q_{S,v} = \sum_{G \in \Omega} I_{\{IG(G,S,v)>0\}} \frac{Score(G, D_0, S, v, \delta)}{\sum_{G \in \Omega} Score(G, D_0, S, v, \delta)} \tag{3.13}$$

where v is the candidate that maximizes EIG, and $I_{\{\}}$ is the indicator function. We add v to the set of interventions if $q_{S,v} > \alpha$, and stop if $q_{S,v} \leq \alpha$. Higher values of α will result in a smaller intervention batch. Setting probability threshold α to ~ 0 (i.e., stopping when the probability of at least some information gain is near 0) is equivalent to stopping when expected information gain is near 0.

When the signal-to-noise ratio in the historic data is low, there is greater weight on graphs where the most optimal intervention candidates have no information gain. This leads to the triggering of the stopping criteria with a smaller batch of interventions. Thus when there is more uncertainty in the data, the procedure avoids the risk of wasteful use of interventions, instead favoring running the experiment with a smaller batch and then relying on the new experimental data to evaluate unused interventions.

3.3.3 Inference of causal network from data acquired post-intervention

The next step in the investigation is to apply the selected interventions, and collect a new dataset D. The new dataset updates Equation 3.7 as

$$\pi(G|S, D_0, D, \delta) \propto p(D|S, G)\pi(G|D_0, \delta) \tag{3.14}$$

The updated edge probabilities are obtained as in Equations 3.8 and 3.9. They balance the canonical representation of the signaling with the condition-specific signaling behavior quantified under the condition of interest, after the interventions. A large deviation of the posterior edge probability from the prior in δ indicates network rewiring or deregulation.

The active learning procedure is iterative in nature, in that the results of the intervention experiment can be viewed as a new instance of historic data. They can inform the selection of new interventions by substituting $\pi(G|S, D_0, D, \delta)$ in Equations 3.11 and 3.12, and repeating the overall procedure.

3.3.4 Implementation and computational complexity

The proposed strategy is implemented in an open-source R package *bninfo*, available on Github. The edge-wise prior is constructed as a data frame in R, either manually or through an interface to the API of KEGG provided by *bninfo*. Historic data is represented as a data frame and pre-selected interventions S as an array. The package *bninfo* implements the algorithms for converting a DAG and a edge-wise to PDAG, calculating the expected gain, and selecting the optimal batch of interventions. The Bayesian network structure learning is performed with the existing R package *bnlearn* [28].

The main scalability bottleneck in the proposed strategy is the selection of interventions in the while loop in Algorithm 3. The complexity of calculating the expected information gain for a single intervention (Algorithm 2 and line 7 in Algorithm 3) is in the order of the number of edges in the input DAG. However, the interventions in a batch are selected sequentially (i.e., the selection of the jth member of the batch depends on the $j-1$ previously selected interventions). Therefore, the selection of j candidate interventions requires up to $j!$ calculations of the expected gain. The running time can be reduced by parallelization of Algorithm 2, or by limiting the number of candidate interventions. Moreover, sampling Ω from

a distribution of DAGs has well-known scalability challenges in Bayesian literature. In the worst case, the computational time scales exponentially with the number of proteins. Bayesian bootstrap sampling can be split among nodes on a cluster, and sped up by parallelization. For the datasets described below, the generation of intervention batches took 70 minutes (17 protein DREAM4 network with 14 candidate interventions and 500 sampled graphs) and 20 minutes (11 protein T-cell network with 5 candidate interventions and 500 sampled graphs) on a 16 node cluster.

3.3.5 Metrics used for performance evaluation

We evaluated the proposed strategy using datasets containing some notion of "ground truth", i.e. situations where the true structure of the causal graph is known. We used the proposed active learning strategy to determine an optimal intervention batch S. To evaluate the performance of S, we considered the data D that would be experimentally acquired with the selected interventions. We then inferred causal networks from D, and derived posterior edge probabilities as described in Section 3.3.3. Finally, we evaluated whether S can lead to network inference that correctly detects edges in the "ground truth" graph.

We evaluated the performance of edge detection using two metrics. The first is the *true positive rate of edge detection*, i.e. the proportion of correctly detected edges among the edges in the ground truth network [52]. Our cutoff for edge's detection is presence and orientation probabilities greater than uninformative prior probabilities described in Equation 3.6. The second metric is the L_1 edge error, which quantifies the overall probability of prediction error, i.e. the probability of either

discovering a false edge or missing a true edge [15, 18]. Given the set of interventions S and the ground truth network, the L_1 edge detection error is defined as

$$\begin{aligned} L_1(G,S) &= \sum_{u,v} \overrightarrow{I}_{uv}(G)(1-\overrightarrow{\pi}_{uv|D,S}) \\ &+ (1-\overrightarrow{I}_{uv}(G))(\overrightarrow{\pi}_{uv|D,S}) \\ &+ (1-\overline{I}_{uv}(G))(\overline{\pi}_{uv|D,S}) \end{aligned} \tag{3.15}$$

where $\overrightarrow{I}_{uv}(G)$ and $\overline{I}_{uv}(G)$ are as in Equation 3.9.

3.4 Data

There are currently no publicly available datasets that implement active learning approach to causal network inference. We therefore use two publicly available datasets, adapted to provide a measure of "ground truth".

3.4.1 DREAM4 Network

"Ground truth" network The 17-node network in Figure 3.3A was used in the DREAM4 Predictive Signaling Network Challenge [23]. The network contains canonical pathways downstream of four receptors (dark grey nodes in Figure 3.3A): two inflammatory (TNFa, IL1a), one insulin (IGF-I), and one growth factor receptor (TGFa). We use this network to evaluate the relative advantages of active learning, and the importance of the use of prior information.

Prior information regarding the network structure We used as the background information the fact that TNFa, IL1a, IGF-I and TGFa are receptors, i.e. proteins that receive signals from the environment, and activate downstream proteins. This presents causal information, in that the remaining proteins in the pathway are downstream of the receptors.

Historic data We use the DREAM4 challenge as a historic dataset. The challenge provided antibody-based measurements (sandwich immunoassays with the Luminex

xMAP platform) from hepatocellular carcinoma cell lines (HepG2), which quantified the activity levels of the signaling proteins at bulk (i.e., non-single-cell) resolution. The dataset is comprised of samples with 5 stimulus conditions, namely no stimulus, stimulus on TNFa, on IL1a, on IGF1, and on TGFa. The dataset also includes 5 intervention conditions, namely no inhibition, inhibition on ikk, on mek12, on pi3k, and on p38. In total there were 25 samples, one for each stimulus and intervention pair.

We processed the historic dataset as follows. We imputed missing values using a neural network model that predicted missing values of a protein given the values of the protein's neighbors in the ground truth network. Several proteins from the pathway (map3k7, ras, map3k1, and mkk4) were not quantified in the historic dataset. Approaches exist for learning Bayesian network structure with hidden variables [62, 63], but these are beyond the scope of this manuscript. So to eliminate this artifact we applied a model that predicts a protein's values given the values of its parents in the model and common biochemical assumptions on signaling dynamics [64], and used the model to predict the values for the hidden nodes. The publicly available challenge data is normalized to the 0-1 range, we then discretized the quantification values to binary on/off variables using .5 as a cutoff.

Candidate interventions All the non-receptor nodes in the network were considered as candidates for interventions. Due to the small number of samples and inhibitions, it was not possible to set aside a portion of this dataset for performance evaluation. Therefore, we fit a causal network model to the challenge data and the ground truth network, and used the model fit to generate synthetic post-intervention datasets. The simulation mimicked the design of the challenge data, in that it contained one biological sample for each intervention. We then evaluated the performance of the selected interventions on these synthetic datasets.

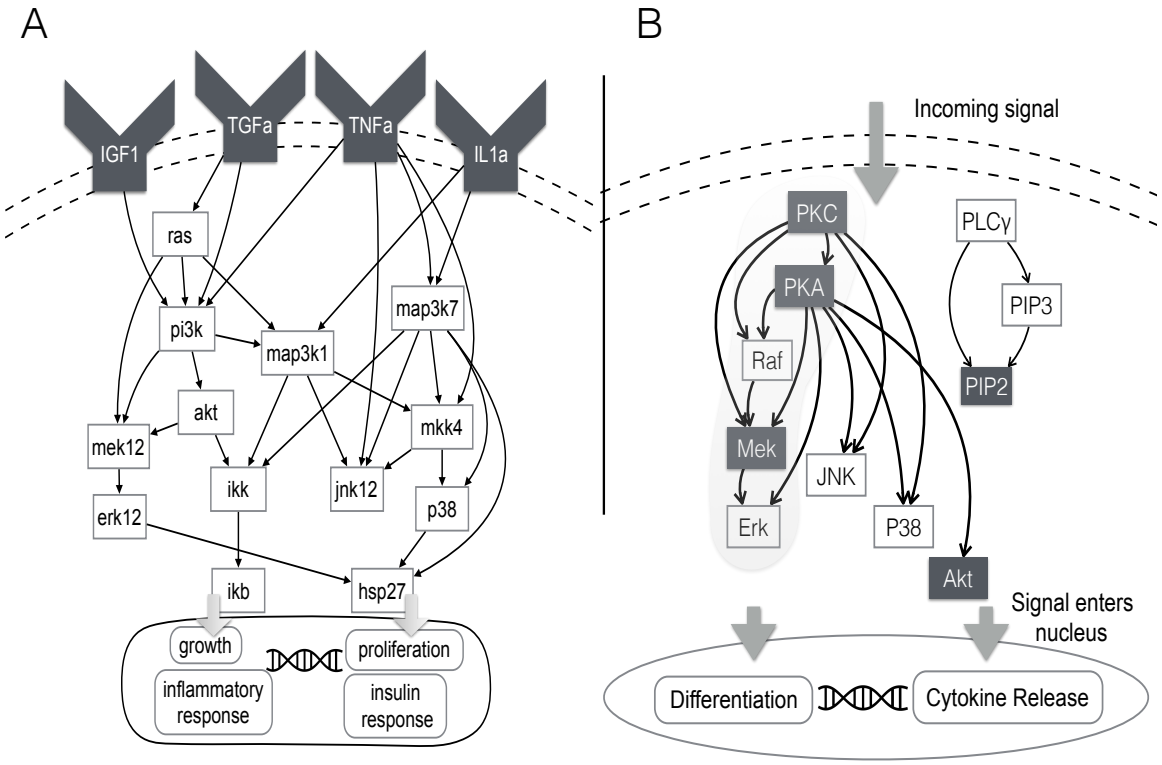

Fig. 3.3. The "ground truth" networks in the experimental datasets. A: The DREAM4 Predictive Signaling Network Challenge network. Dark nodes indicate receptors. B: Native T-cell signaling network in response to antigen. The edges in the PKC $\rightarrow$ Raf $\rightarrow$ Mek $\rightarrow$ Erk cascade, and the edge PKA $\rightarrow$ Raf belong to the canonical MAPK pathway. Dark nodes indicate targets of experimental interventions.

3.4.2 Flow Cytometry Measurements of T-cell Signaling

"Ground truth" network The network in Figure 3.3B contains 11 phosphoproteins and phospholipids involved in the native CD4+ T-cell signaling response to antigen, and their canonical edges. The network was used to validate causal inference from an experimental dataset [11], and has been subsequently used as a benchmark in multiple causal inference studies [4, 19].

Prior information regarding the network structure We used as the background knowledge the edges in the canonical MAPK pathway. Although CD4+ T-cell signaling has been extensively studied, we assumed that no prior information is available regarding the remaining edges. This assumption allowed us to compare the case of a minimally informative prior to the case of a uninformative prior, and focus performance evaluation on detection edges that were not addressed in the background knowledge.

Historic data The experimental dataset in [11] contains single cell fluorescence-based quantifications of 11 phosphoproteins and phospholipids in human primary naive CD4+ T-cells. These analytes are downstream of the receptors CD3 and CD28 that provide co-stimulatory signals required for T-cell activation. We used as the "historic data" a portion of this dataset that was acquired without any targeted interventions. This portion of the dataset contained 11672 cells.

Candidate interventions The dataset in [11] also contained single cell quantifications, acquired after activating or inhibiting five signaling proteins (dark nodes in Figure 3.3B). These five proteins were considered as candidates for interventions in this manuscript. The post-intervention experimental datasets were used to compare the information gain projected in our approach with the actual information gain after the interventions.

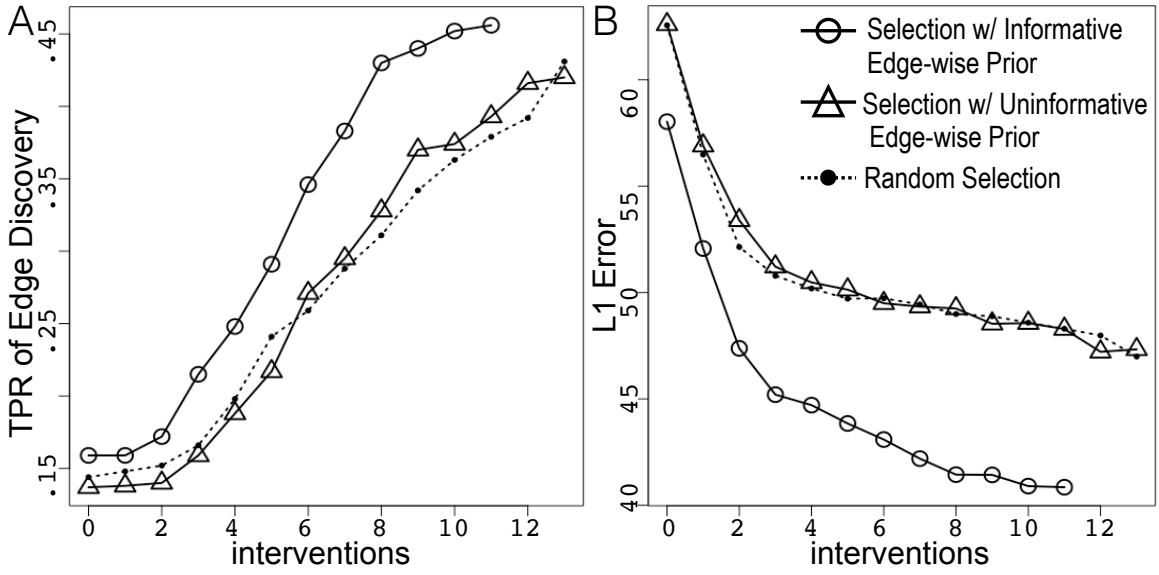

Fig. 3.4. Performance of the proposed strategy on the DREAM4 dataset. Dotted line: uninformative edge-wise priors, randomly selected interventions. Solid line with triangles: uninformative edge-wise priors, interventions selected with active learning. Solid line with circles: informative edge-wise priors, interventions selected active learning. A: True Positive Rate (TPR) of detecting ground truth edges. B: L_1 error of detecting ground truth edges.

3.5 Results

3.5.1 Informative prior edge probabilities reduced the required number of interventions in the DREAM4 dataset

In the DREAM4 dataset, we compared (1) random ordering of interventions, and (2) proposed active learning strategy with uninformative prior probabilities of edge presence (Equation 3.6), and (3) the same active learning strategy, but with informative priors. The informative priors encoded the fact that the receptor nodes have upstream positions in the network. All the receptor-originating edges were assigned the prior orientation probability of ~ 1, all receptor-terminating edges were assigned the presence probability of ~ 0, and the remaining edges were assigned the uniformed prior presence probability in Equation 3.6. All the non-receptor nodes in the network were candidate targets for interventions. In order to exhaust the information in the prior and historic data, the active learning approach with both priors used the most liberal stopping criteria for growing a batch. It only stopped adding interventions when all the additional candidates has expected information gain of ~ 0.

Figure 3.4 summarizes the results. With uninformative edge-wise priors, random selection of interventions performed similarly to the proposed active learning in both True Positive Rate of edge detection and L_1 loss. Both metrics depend on correct detection of edge presence as well as edge orientation. The contribution to edge detection of the added samples provided by each intervention experiment overshadowed the selection strategies prioritization of interventions that better resolve edge orientation.

At the same time, the results demonstrate the efficiency gain in selecting the interventions, brought by encoding the knowledge of receptor identities into the prior. For example, while with the uninformative prior the true positive rate of more than .35 could be achieved with on average 9 interventions, the informative prior only required on average 6 interventions. Table 3.1 shows the specific interventions that

Prior	Intervention targets selected by active learning
Informative	(1) hsp27, (2) mek12, (3) map3k1, (4) jnk12, (5) pi3k, (6) mkk4, (7) ikk, (8) akt, (9) p38, (10) erk12, (11) ikb
Uninformative	(1) jnk12, (2) hsp27, (3) mkk4, (4) mek12, (5) pi3k, (6) map3k1, (7) map3k7, (8) ras, (9) ikk, (10) akt, (11) ikb, (12) p38, (13) erk12

Table 3.1.
Intervention targets selected by active learning in the DREAM4 dataset. The informative prior edge probabilities required a smaller intervention batch.

were selected at a conservative cutoff ($q_{S,v} \leq 0.01$). The results indicate that informative edge-wise priors are important for such bulk experiments with a small number of replicate samples. The prior knowledge removed uncertainty in edge presence, increasing the contribution of improved detection of edge orientation to overall performance.

3.5.2 The ordering of T-cell interventions by active learning matched their contribution to causal inference

As above, for the T-cell dataset we compared (1) random ordering of interventions, (2) proposed active learning strategy with uninformative prior probabilities of edge presence (Equation 3.6), and (3) the same active learning strategy, but with informative priors. In the latter case we assumed the prior knowledge of the canonical MAPK pathway, and assigned a high prior probability (~ 1) to the edges in the PKC $\rightarrow$ Raf $\rightarrow$ Mek $\rightarrow$ Erk cascade, and to the edge PKA $\rightarrow$ Raf. The remaining potential edges were assigned the uninformative prior probability of both presence and orientation. We then considered the five interventions in [11] as the set of candidate interventions.

Figure 3.5 summarizes the results. Selection with an uninformative edge-wise prior did not outperform random selection of interventions in terms of True Positive Rate of detecting edges. However, it had a smaller L_1 error for the first three selected interventions. With this experiment, the intervention datasets served not just to resolve causality, but to improve edge detection by adding variation in signaling activity not present is preceding datasets. As with the DREAM4 data, this lead to performance gains due to improved edge detection overshadowed gains owed to the selection strategy.

In contrast, active learning with the edge-wise prior encoding the MAPK edges outperformed random selection both in terms of greater True Positive Rate, and smaller L_1 error. Table 3.2 shows the specific interventions that were selected at a conservative cutoff ($q_{S,v} \leq 0.01$). For example, while with the uninformative prior the true positive rate of .75 could be achieved with on average 5 interventions, the informative prior only required on average 3 interventions.

The network structure provides some insight into the role of informative edge-wise priors in improving the performance. Since the orientation probability of an edge depends on the orientation probability of its neighbors, the edge-wise prior reduced error by reducing uncertainty in the orientation of edges neighboring the MAPK edges. In addition, the edge-wise prior enabled the causal inference procedure to down-weight graphs where the MAPK edges were not present or had the wrong orientation, increasing sensitivity. Finally, additional causal information encoded in the prior made interventions on Mek and Akt interventions less useful, enabling the stopping criteria to eliminate them from the batch.

3.6 Discussion

Our results showed that an active learning strategy, combined with informative priors, is the most effective at suggesting the smallest batch of target interventions for

Prior	Intervention targets selected by active learning
Informative	(1) PKA, (2) PKC, (3) PIP2
Uninformative	(1) PKA, (2) PIP2, (3) PKC, (4) Akt, (5) Mek

Table 3.2.
Intervention targets selected by active learning in the T-cell dataset. The use of an informative edge-wise prior eliminates two interventions from the batch.

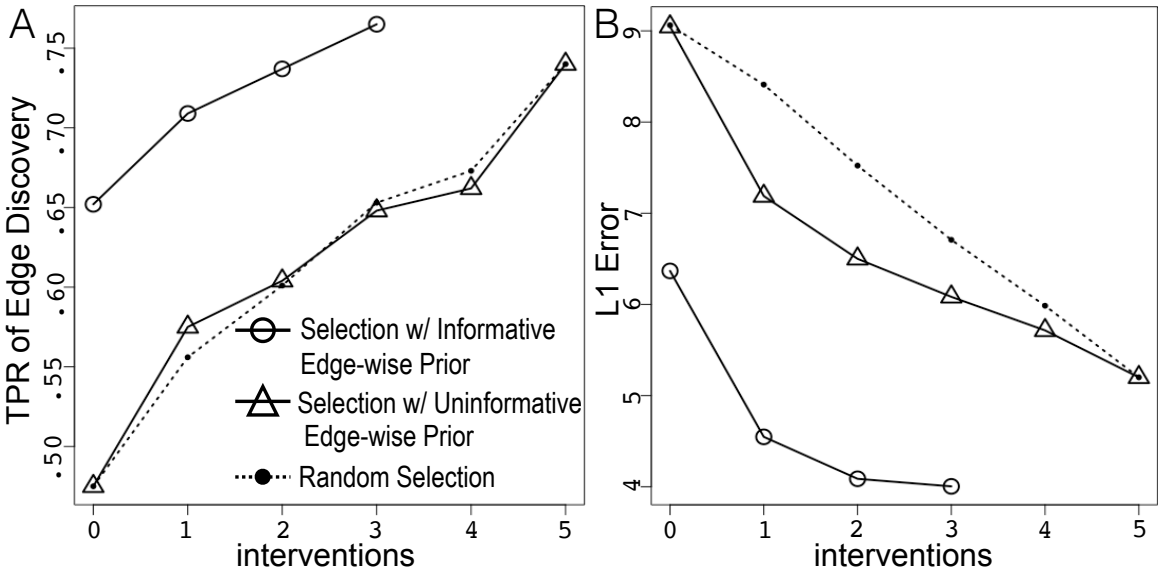

Fig. 3.5. Performance of the proposed strategy on the T-cell signaling dataset. Lines and panels are as in Figure 3.4.

inference of causal networks. It optimizes the causal information in the intervention experiments, while controlling the experiment time and cost.

The background information comes in the form of prior knowledge on the presence and orientation of edges in the system available, e.g., in pathway databases such as KEGG, as well as from historic datasets available, e.g. from repositories such as Cytobank. The proposed strategy can be used with any level of prior information or uncertainty. When prior information and historic data indicate an intervention is potentially wasteful, it will tend to omit it from the batch, electing rather to reserving it for potential use in future experiments.

The active learning strategy in this manuscript suggests interventions based on the prior information and on the topology of the network. In practice, however, other factors can affect the utility of an intervention. For example, small molecule inhibitors vary both in cost and efficacy. Their inhibitory effects may only occur with some probability, and may also have off-target effects. The proposed framework can be easily extended to incorporate this type of "soft intervention" [4], as well as cost

considerations into the expected information gain. Alternatively, it can produce a batch of interventions with a fixed pre-specified number of targets, while selecting the targets with the most expected causal information gain.

The proposed methodology relies on signaling network modeling by means of direct acyclic graphs. In practice, however, cell signaling often displays feedback loops. This can be addressed by refining the biological interpretation of the graph structures. In particular, cycles in signaling often involve regulatory feedback loops that involve transcription. In this case, an activation of the signaling pathways causes transcription, which then results in the translation of new signaling proteins which then change the initial signaling pathway. Since the time scale of signaling is in seconds and minutes, and the time scale of transcription is in minutes and hours, collecting the data at an appropriate time point can help resolve the confounding between the initial causal effect of the signaling, and the feedback.

This work addressed both the inference of causal networks from bulk experiments and single cell experiments. Inferring an edges orientation depends of first detecting its presence, and since many edges between the proteins can potentially exist, bulk experiments often lack replicates to confidently detect the edges. Therefore, in bulk experiments it is often useful to add interventions not only to improve the detection of edge orientation, but also to improve the detection of edge presence through increased sample size.

In contrast, single cell experiments characterize protein signaling activity in thousands to millions of individual cells, and increase our confidence in the inferred edge. If the historic data was collected under a minimal number of conditions, intervention data may resolve both edge orientation and presence by virtue of adding variation in signaling activity. Moreover, single-cell experiments do not eliminate the need for true biological replication. The proposed approach can be extended to population level inference by modeling subjects as additional nodes in the network [65].

Overall, we believe that the proposed strategy is an important step towards an informed experimental design for inference of causal networks, and advocate its practical use.

4. BAYESIAN INFERENCE OF KINETIC PARAMETERS FROM SINGLE-CELL DATA

4.1 Introduction

Cell signal transduction describes intracellular protein regulatory networks that determine how a cell reacts to its environment [1]. Understanding the structure and dynamics of the biochemical interactions that comprise cell signaling pathways is an important task, leading to improved insight into disease states and increased ability to intervene towards therapeutic ends, for instance, in diseases such as cancer.

The computational systems biology approach to the study of cell signaling is to build a *kinetic model*, comprised of a set of biochemical reactions that capture the behavior of a signal transduction response of interest [7]. The utility of such a model is that it can predict the outcomes of interventions to the system that are not seen under normal signaling conditions, eg. a drug intervention. Such predictions can be used to design follow up experiments that validate those predictions. Computing standards, such as systems biology markup language (SBML) [66], and public databases of published models such as Biomodels.org [67] enable the digital sharing of kinetic models so that investigators can reproduce, validate, and build from the work of others.

Kinetic models have *rate parameters* (constants) that determine the rates of reaction. We use experiments to estimate rate parameters. Traditionally, kinetic models are built using the "bottom-up" approach, where rate parameters are inferred using information from multiple purified in vitro enzyme kinetics experiments. The advent of high-throughput omics technologies motivates the "top-down" approach, which seeks to do system-wide inference of model parameters using a single experi-

mental design [68]. The attractiveness of the top-down approach lies in that it is a data-driven approach to elucidating the mechanics underlying whole systems.

In the context of cell signaling, proteomics technologies provide information on the activity of signaling proteins. Targeted proteomics platforms such as mass cytometry (CyTOF) that collect single cell level quantifications have proven a valuable tool in system-wide inference of *causal Bayesian network* models of signaling [37]. These experiments involve many thousands of single cell level quantifications of the enzymatic activity, there is adequate statistical power to infer the presence and orientation of causal regulatory relationships between proteins. However, each cell measurement is only a single time point "snapshot" of signaling taken from the full time course of the cell's signaling behavior. So it remains unclear how to use this *single cell snapshot data* to learn about the biochemical kinetics underlying these relationships.

This manuscript demonstrates how to quantify causal influence in causal Bayesian network model of signal transduction from single cell snapshot data. Our approach uses classical chemical kinetic modeling (eg. mass action or Michaelis-Menton kinetics) to quantify the relationships in a causal Bayesian network in terms of kinetic model rate parameters. We then estimate those rate parameters from data.

In the Background section we introduce key concepts in quantitative modeling of signal transduction, the state-of-the-art in inferring those models from snapshot data, and the challenge of connecting the results of inference to reaction kinetics. The Methods section follows with details of how to construct a causal Bayesian network model that solves this problem. In the Evaluation section we validate our methods using synthetic data to reproduce kinetic parameters from a highly-cited mathematical model. We also replicate the results of a published machine learning approach to fitting a curve between two signaling proteins quantified in CyTOF data, and show that the proposed approach produces additional insights into the kinetics of the system.

4.2 Background

4.2.1 Kinetic models of cell signal transduction

Kinetic models of cell signaling model specify a set of species (signaling protein isoforms), biochemical reactions involving those species, rate parameters that describe those reactions, and rate laws that determine the rate of each reaction's occurrence in terms of the rate parameters. The species in kinetic models are signaling proteins in various states of enzymatic activity, the simplest case being the states "active" and "inactive". Rate laws vary in complexity, generally less complexity means fewer parameters. Typically, the modeler assumes a rate law that is as simple as possible while still being faithful to observed signaling behavior. With these components, the kinetic model describes the evolution of the abundance of each species. An example is the canonical signaling model of the phosphorylation cascade, where a protein is in an "active" state if it has been phosphorylated, and a phosphorylated protein acts as an enzyme catalyzing the phosphorylation of another protein. Each phosphorylation reaction might be modeled with a *Michaelis–Menten* rate law, which assumes the rate of phosphorylation of a protein depends on abundance of the enzyme, abundance of the unphosphorylated substrate, and rate parameters corresponding to the binding and unbinding of the enzyme to the substrate and the rate the bound compound picks up a phosphate group. Signal flows through this cascade of phosphorylations. A kinetic model of such a cascade would model each phosphorylation with a separate reaction and reaction rate.

Analysis of signal transduction is often focused on *steady state* signaling, meaning the state of the system after an initial period of transience as the external signal is processed. Generally, the signaling response occurs after the system reaches steady-state, eg. transcription of a gene [69]. In practice steady states are not permanent, because extrinsic factora, including the signaling response itself, alter the environment of the cell. For example, a signaling response can involve transcription of a gene that is translated into a new signaling protein that in turn affects the signal-

ing pathway [70]. In such cases it is more appropriate to say the analysis targets a *quasi*-steady state. In this work, we focus on signaling systems with quasi-steady states that are stable, meaning after steady state is reached the abundance of protein species doesn't change in time. Part of the contribution of this work is motivating the analysis of a signaling mechanism's steady state behavior when preparing to model it with single cell snapshot data.

A central goal of computational systems biology is to estimate rate parameters. Typically this is done with low throughput time course data, because rate parameters themselves are not identifiable from "snapshots" taken at a single time point ([71, 72]). The contribution of this work is a method for inferring identifiable functions of rate parameters from high throughput single cell snapshot data, and using the results of this inference to inform direct inference of rate parameters from low throughput time course data.

4.2.2 Single cell data, cell variability, and stochastic modeling

This work focuses on single time point cell-level "snapshots" of signal transduction produced by single cell proteomic experiments. In these experiments cells are exposed to a signaling stimulus, then measurement platforms collect cell-level snapshots of intracellular signaling activity. In single cell flow cytometry, cells are chemically fixed, intracellular signaling proteins are tagged with fluorescently-labeled antibodies, and the cytometer records the antibodies' fluorescent signals in individual cells, each recording reflecting the relative abundance of signaling proteins in different states of enzymatic activity within an individual cell [42]. Mass cytometry (CyTOF) is an alternative to flow cytometry with higher throughput and more precise quantifications. In CyTOF experiments, intracellular signaling proteins are tagged with heavy-metal isotopes, and the cytometer collects and records the mass-to-charge ratio of the charged isotope tags [43].

One way of building a kinetic model from single cell snapshot data is to use a *deterministic approach* [73]. This approach assumes the species are well mixed in a fluid volume with reactions occurring uniformly through time. Based on this assumption, the abundance of a certain species at a specific time point is modeled with a deterministic formalism such as differential equations. Such approaches model deviations in data from deterministic predictions with statistical error [71, 74, 75].

However, there is a large amount of variability in signaling states between cells, even among populations of genetically identical cells in uniform environmental conditions [72, 76]. Single cell experiments will demonstrate this cell-to-cell variability, in contrast to bulk experiments where quantifications are essentially averages across all the cells in the sample. This variability can be traced to the observation that the biochemical reactions underlying signaling occur with very low numbers of molecules. Such reactions result in unpredictably fluctuating numbers of molecules in individual cells or their compartments within the cell, and thus in different protein abundance across cellular populations. Indeed, attempts to build larger models have suggested that noise can accumulate within signaling pathways [77], such that two cells of the same type same signaling conditions may exhibit entirely different downstream signaling responses. Deterministic methods that try to account of cell variation with statistical error terms may be insufficient.

In contrast, the *stochastic approach* builds a kinetic model that explicitly incorporates cell-to-cell variation in its formulation. It assumes that the event protein substrates in a reaction come in contact is a matter of chance, due to the low abundance of signaling proteins. With this assumption, cell-to-cell variation is a natural result of the inherent physical stochasticity of the system [73, 78].

Our approach models signal transduction as a *continuous time discrete state Markov process* [73, 79]), a stochastic approach where the probability a reaction occurs in the *next* instant depends only on the state of the system at *the current* instant. We then use Markov process theory [73, 80] to derive the steady state probability distri-

bution of possible protein abundances and use this model the cell-to-cell variation in single cell data.

4.2.3 Absolute abundance in single cells

Building systems biology models from proteomics data relies on information about absolute abundance of protein levels in a sample [81, 82]. To date, studies involving single cell snapshot experiments have limited interpretation of the data to relative abundance of protein levels between samples. Statistical methods exist for calibrating quantifications to absolute abundance the domains of both mass spectometry (bulk experiments, not single cell) [83], and flow cytometry [84], suggesting possible extensions to CyTOF. The method proposed here works even if the relationship between quantification values and absolute abundance is ambiguous, and results of the analysis can inform experiments that elucidate that relationship.

4.2.4 Learning causal network models of signaling

Two proteins have a causal relationship if one regulates the other, e.g. two proteins have an enzyme-substrate relationship. The goals of causal modeling in the context of cell signaling are (1) determine whether there is a causal relationship between two proteins, (2) determine the direction of causal influence, and (3) quantify the strength of causal influence [46]. Causal Bayesian network structure inference from protein quantifications have emerged as an ideal machine learning approach for achieving the first and second goal [8, 85], particularly from single cell data [21, 37]. In a causal Bayesian network, nodes are variables representing levels of signaling activity of the proteins. For example, a node can take discrete signaling states such as "active" or "inactive", or a continuous value representing the abundance of the protein in its active state. The model expresses causal relations between nodes with a directed acyclic graph structure (DAG) G. The edge direction in the DAG reflects the causal effects of a change in the signaling state of a parent node on the state of

the child. The challenge of learning a causal DAG representation of a dynamic setting is not a small one, requiring stringent experimental settings [3, 21, 46] that help meet the strong assumptions of causal inference [10].

Each node (protein) in the causal Bayesian network has a conditional probability distribution (CPD) given its parents (regulator proteins). The CPD and its parameters are a probabilistic representation of the regulatory influences of the parents on the child [6]. The joint probability distribution of all the nodes in the network factorize into a product of the CPDs;

$$P(G) = \prod_{j}^{K} P\left(Y_j \mid \mathrm{Pa}(Y_j)\right) \tag{4.1}$$

where P(G) is the joint probability of the K nodes $Y_1 ... Y_K$, and $P\left(Y_j \mid \mathrm{Pa}(Y_j)\right)$ is CPD the jth node Y_j conditional on its parents $\mathrm{Pa}(Y_j)$. Further, each node is conditionally independent of its non-descendent nodes (non-downstream proteins) in the network including its indirect predecessors (upstream proteins, exluding direct regulators).

In this work we focus on the case when causal network structure is known, and the goal is quantifying the causal influence between proteins in network. Information theoretic approaches have provided powerful techniques for cope with this cell variability in single cell data [76, 77, 86, 87]. In this work, we quantify causal influence in terms of a causal Bayesian network CPD. We use a Bayesian approach to infer CPD parameters, using probability to model cellular noise much like information theoretic approaches. However, our proposed model goes a step further by defining model parameters in terms of rate parameters, such that inference on these parameters provides deeper insight into the systems dynamics. Secondly, we specify one sparse multivariate model for an entire system, avoiding the challenges that information theoretic methods face in extending to multivariate settings, as well as enabling us to model noise throughout the system.

Alternative modeling frameworks do explicitly combine network structure with kinetic modeling and rate parameter estimation. For example, Alon et al. [70] mod-

eled signaling with a multilayer perceptron with an activation function based on Michaelis-Menton kinetics, and Terfve et al. [64] built a Boolean network of signaling that assumes Hill kinetics (a generalization of the Michaelis-Menton kinetics). These models validate that causal Bayesian networks can accomplish the same task because the CPDs in Bayesian networks can theoretically represent any set of deterministic functions [10, 88], i.e. these are special deterministic cases of causal Bayesian networks. Our approach combines kinetic modeling with a probabilistic accounting for cell-to-cell variation.

4.2.5 Bayesian inference of kinetic models of signaling

In this work, we propose a Bayesian approach to inferring the parameters of causal Bayesian network CPDs. Here the term *Bayesian* refers to use of Bayesian statistical methods for parameter inference. This meaning is distinct from the term's meaning in "causal *Bayesian* network", where it refers to the factorizability in Equation 4.1. Methods for Bayesian parameter inference in systems biology contrast to optimization methods such as maximum likelihood (see [73, 89] for introductions to Bayesian inference in systems biology, and [74], [71] to compare with optimization approaches). This work focuses on inference with Bayesian Markov chain Monte Carlo (MCMC).

Bayesian MCMC refers to algorithms that take as an input data, a specification of conditional probability of the data given parameters, and *prior* probability distributions on the parameters. These then generate samples from the *posterior* probability of the parameters [90]. Probabilistic programming software that implement MCMC (eg. [91, 92]) allow users to enter data and a set of probability distributions, compile the inputs into a custom MCMC algorithm, then output results in the programming environment.

Prior work has applied biological interpretation only to DAG structure, and has not extended it to the CPD. Rather, investigators have chosen CPD families that fit

software constraints but lacks biological interpretation (eg. Gaussian or multinomial, [8, 85]. To our knowledge, no publicly available Bayesian network software allows for custom CPDs. In contrast, we use the stochastic approach to kinetic modeling to specify CPDs that explicitly model cell-to-cell variability and rate parameters. We then infer the parameters of the CPD's using probabilistic programming. This closes the gap between saving time and avoiding human error by implementing parameter inference with publicly available software tools, and having a probability model that is specific to the problem domain.

4.3 Methods

4.3.1 Defining kinetic rate laws

We model a molecule of signaling protein such that its physical configurations map to two states; enzymatically "active", and enzymatically "inactive". When there are more than two physical configurations, one configuration is labeled "active", the others are collectively labeled "inactive".

Notation. The decomposability (Equation 4.1) of a causal Bayesian network model of signaling enables its breakdown into individual sets comprised of an individual protein node and its parents in the network. We make use of this parsimony to model each node separately. In each node-parents set, y refers to a node, x_1, ..., x_p to its p parents in the DAG. Y denotes the absolute abundance of active y, Y^o denotes the absolute abundance of inactive y. N_y denotes the total abundance of y. N_y is a *conservation rule* for y, meaning its value is "conserved" even as Y and $Y^o = N_y - Y$ vary in time [70,73]. X_j denotes the abundance of active x_j. Y and N_y take positive-real values. For now assume the units of absolute abundance reflect molar concentration, the standard for most SBML models [66], though we generalize the units in a later section.

By adopting common assumptions of the dynamics of signaling, we can quantify the causal influence of x_j on y in terms of kinetic rate parameters. In this manuscript,

we focus on the two most common kinetic assumptions used in signaling models; mass action kinetics and Michaelis-Menton kinetics.

Mass action rate law. Suppose y only has one regulator (i.e. parent in the DAG) x. In a mass action kinetic model, the activation of y by x is represented with the chemical reactions

$$X + Y^o \xrightarrow{v} X + Y \tag{4.2}$$

$$Y \xrightarrow{\alpha} Y^o \tag{4.3}$$

This system has two reactions, each with a rate parameter; v is the rate parameter for the activation reaction (reaction 4.2), and α is the rate parameter for the deactivation reaction (reaction 4.3). When reactions occur continuously in time (deterministic model of the system) the overall rate of change of Y is:

$$\frac{dY}{dt} = vXY^o - \alpha Y \tag{4.4}$$

The rate of activation from inactive to active state is the function:

$$\text{rate of activation} = vXY^o \tag{4.5}$$

Michaelis-Menten first order rate lawl. Activation by X in the Michaelis-Menton model expands the mass action model with a binding reaction;

$$X + Y^o \underset{k_{\text{diss}}}{\overset{k_{\text{bind}}}{\rightleftharpoons}} [XY^o] \xrightarrow{v} X + Y \tag{4.6}$$

$$Y \xrightarrow{\alpha} Y^o \tag{4.7}$$

The reaction equation 4.6 is shorthand combining 3 reactions; a binding reaction with rate parameter k_{bind}, a reaction for the disassociation of the bound compound with rate parameter k_{diss}, and an activation reaction with rate parameter v This adds two additional rate parameters (k_{bind} and k_{diss}) compared to mass action kinetics. By adding binding into the model, the rate of activation changes from the vXY^o in the mass action model to

$$\text{rate of activation} = \frac{vXY^o}{K + Y^o} \tag{4.8}$$

K is the Michaelis-Menton constant and is a function of the rate parameters;

$$K = \frac{v + k_{\text{diss}}}{k_{\text{bind}}} \tag{4.9}$$

and is interpreted as the concentration of Y_o at which the rate of activation is half of the maximum rate possible.

In this work we employ the *first-order kinetic assumption*, a common simplifying assumption for Michaelis-Menton kinetic models. This assumption simplifies Equation 4.8 to:

$$\text{rate of activation} = \frac{vXY^o}{K} = \tilde{v}XY^o \tag{4.10}$$

where $\tilde{v} = \frac{v}{K}$. This assumption is appropriate when Y^o remains very low (i.e. $Y^o << K$), which is typically the case in context of cell signaling [70]. The rate of change of Y is therefore:

$$\frac{dY}{dt} = \tilde{v}XY^o - \alpha Y \tag{4.11}$$

In both the mass action and Michaelis-Menton models, we model deactivation of y with a single rate parameter α. Note that deactivation can be modeled more explicitly with a deactivating enzyme, such as a phosphotase. In this case, the deactivating enzyme is a separate parent node (i.e. regulator).

4.3.2 Modeling the regulation of signaling for each node

Steady state assumption. A fundamental assumption of our overall approach is that we model the system at a stable steady state. Single cell snapshot data can only quantify signaling activity in a cell at a single time-point. Theoretically it is possible to apply a signaling stimulus, and label each cell with the elapsed post-stimulus time when the cell is quantified. But in practice there is no way to assure the recorded time point is the "true" post stimulus time point during the transient signaling response. By choosing a time point past the time the system achieves

steady state, we comfortably assume variation between cells is due to cell-to-cell heterogeneity, and not to transient cell response.

We define a *regulatory function* g that maps the states x_1, ..., x_p to the state of y. g is effectively a nonlinear regression of Y on $X_1...X_p$ and is defined in terms of a set of nonlinear regression parameters. While the rate parameters are not identifiable from steady-state snapshot data, the regression parameters are functions of the rate parameters that are identifiable at steady state. The choice of the g depends on a priori knowledge on how y is regulated by its parents. We describe three types of regulatory functions below.

2-state activation. The steady state solutions of Equations 4.4 and 4.11 are;

$$\frac{Y}{N_y} = \frac{\beta X}{1 + \beta X} \tag{4.12}$$

The regression parameter β is a function of the rate parameters given by the kinetic assumption. Under the mass action kinetic assumption in Equation 4.4:

$$\beta = \frac{v}{\alpha} \tag{4.13}$$

Under the Michaelis-Menton assumption in Equation 4.11:

$$\beta = \frac{\tilde{v}}{\alpha} \tag{4.14}$$

The regression parameter β has the straight-forward interpretation as a quantification of the strength of influence of x on y.

Regardless of whether using the mass action or Michaelis-Menton assumption, the regulatory function generalizes to the case of multiple parents;

$$\frac{Y}{N_y} = g_1(\sum_j \beta_j X_j) \tag{4.15}$$

$$g_1(u) = \frac{u}{1 + u} \tag{4.16}$$

The sign of β_j is positive if the j^{th} parent is an activator, negative if it is an inhibitor.

3-state activation. A common signaling mechanism is double-phosphorylation, where a protein reacts with an enzyme twice before it becomes active. In this case,

N_y is the sum of concentrations of 3 states; Y^o, Y^p, and Y, where Y^p is concentration of y in an intermediate state, such as having only one single phosphate group when two are required to be active. In the case of mass action kinetics

$$\begin{aligned} X + Y^o &\xrightarrow{v} X + Y^p \\ X + Y^p &\xrightarrow{v} X + Y \\ Y &\xrightarrow{\alpha} Y^p \\ Y^p &\xrightarrow{\alpha} Y^o \end{aligned} \tag{4.17}$$

Under the assumption that the kinetic rate parameters (including the binding rates in the case of the Michaelis-Menton assumption) are the same for both the $Y^o \leftrightarrow Y^p$ and $Y^p \leftrightarrow Y$ state transitions, the regression parameter β retains the same definition as before, and the steady-sate solution in Equation 4.16 becomes:

$$\frac{Y}{N_y} = g_2(\sum_j \beta_j X_i) \tag{4.18}$$

$$g_2(u) = \frac{u^2}{1 + u + u^2} \tag{4.19}$$

Feedback Loops. Feedback loops are typically modeled with cycles in directed network visualizations of signaling, a graphical representation that seemingly violates the acyclicity assumption of the causal Bayesian network. However, under the stable steady-state assumption, the Markov condition still holds. We demonstrate this with the example in Figure 4.1. In Figure 4.1 signal is passed from x to y_1 to y_2 to z, but y_2 also inhibits y_1 creating a negative feedback loop. Let Y_1^{off} and Y_1^{on} represent respectively the inactive and active concentrations for y_1, likewise let Y_2^{off} and Y_2^{on} be defined the same way for y_2. Let $N_1 = Y_1^{\text{off}} + Y_1^{\text{on}}$ and $N_2 = Y_2^{\text{off}} + Y_2^{\text{on}}$. Assume the following kinetic model

$$\frac{Y_1^{\text{on}}}{dt} = vXY_1^{\text{off}} - \delta Y_1^{\text{on}} Y_2^{\text{on}} - \alpha Y_1^{\text{on}} \tag{4.20}$$

$$\frac{Y_2^{\text{on}}}{dt} = vY_1^{\text{on}} Y_2^{\text{off}} - \alpha Y_2^{\text{on}} \tag{4.21}$$

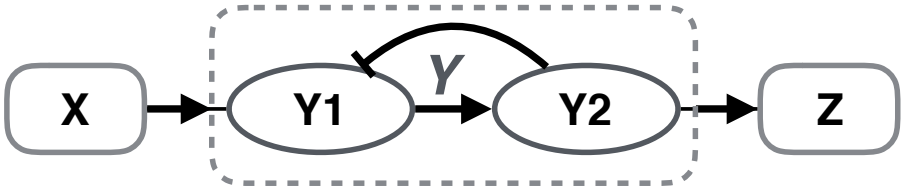

Fig. 4.1. Example of a negative feedback loop. Feedback loop diagrams seem to violate the no-cycle constraint of a causal network. However, if there is stable steady state the causal network model can still be applied.

The term $\delta Y_1^{\text{on}} Y_2^{\text{on}}$ in Equation 4.20 corresponds to y_2's inhibition of y_1. The steady state of y_2 is as in previous cases

$$\frac{Y_2^{\text{on}}}{N_2} = g_1(\beta Y_1^{\text{on}}) \tag{4.22}$$

Solving for the steady state of Y_1 involves setting Equation 4.20 equal to 0, plugging in the value for Y_2^{on} from Equation 4.22, then solving for $\frac{Y_1^{\text{on}}}{N_1}$

$$\begin{aligned} \frac{Y_1^{\text{on}}}{N_1} &= \frac{\beta X}{\beta X + \frac{\delta}{\alpha} Y_2^{\text{on}} + 1} \\ &= \frac{\beta X}{\beta X + \frac{\delta}{\alpha} N_2 g_1(\beta Y_1^{\text{on}}) + 1} \qquad (4.23) \\ &\leq g_1(\beta X) \qquad (4.24) \end{aligned}$$

Solving Equation 4.23 for $\frac{Y_1^{\text{on}}}{N_1}$ yields a regulatory function that depends on X, N_2, β and δ, and not on any concentration values for the downstream node y_2. The acyclicity assumption still applies because y_1's regulatory function depends not on the concentration values of downstream nodes, but rather on rate parameters involved in downstream reactions.

Explicitly finding the regulatory function for y_1, the node directly regulated by the loop, is non-trivial in the case of most loops. This requires solving Equation 4.23 for Y_1^{on}, i.e. solving a polynomial expression that becomes increasingly complex with the size of the loop. Moreover the regulatory function depends on the size of

the loop and the number of associated rate parameters, preventing a generalized outcome for g as in Equations 4.16 and 4.19.

We address this by working establishing with bounds the regulatory function. In the case of the negative feedback loop in Figure 4.1, $g_1(\beta X)$ forms a lower bound on Y_1^{on}'s regulatory function, as shown in Equation 4.24. Note that if the rate parameter for inhibition δ were small, then g_1 would be a good approximation.

We also propose an alternative approach of simplifying the problem by collapsing the loop into a single *super-node*. For example Figure 4.1 demonstrates the collapse of y_1 and y_2 into a super-node y. Let the concentration $Y = Y_1^{\text{on}} + Y_2^{\text{off}} + Y_2^{\text{on}}$. Then the change in Y concentration over time is given by

$$\frac{dY}{dt} = vXY_1^{\text{off}} - \alpha Y_1^{\text{on}} \tag{4.25}$$

Substituting N_1 and N_2 into Equation 4.25 and finding the steady state solution yields

$$\frac{Y}{N_1 + N_2} = \frac{\beta X + \frac{N_2}{N_1+N_2}}{\beta X + 1} \tag{4.26}$$

This approach of collapsing feedback-loops into super-nodes and plugging in values for the conservation rules generalizes well to other types of feedback loops.

Generalized regulatory functions. In this subsection we described a regulatory function g that determines how a protein's state is regulated by the states of its parents in the network. The regulatory function is identified by solving for the ratio of active concentration to total concentration at steady state. We outlined two cases g_1 and g_2 that are derived directly from common signaling kinetics rate laws. We addressed how a bespoke g can be derived for a specific case of a feedback loop, a common functional motif in biochemical networks. An in-depth examination of how to derive g for all possible motifs, loops, parameter sets, and signaling interactions is beyond the scope of this manuscript. Below, for a given protein we refer to the protein's regulatory function simply as g, and assume the appropriate form of g has been established.

4.3.3 Deriving the conditional probability distribution for a node

We now use the generalized regulatory function to define the CPD for each node in the network.

Experimental setting. Experimental data is acquired by applying signaling stimulus to prepared samples, waiting through a period of transient dynamics as the cell initially reacts to the signal, chemically fixing the cells when it is believed the signaling response has reached a steady state, then finally permeabilizing and staining the cells before submitting them to the CyTOF for quantification. We infer the coefficient parameters from experimental data. Each cell level quantification is treated as a replicate measurement of protein abundance for each targeted protein.

Constructing the data model. Let Y_i and $N_{y,i}$ denote active state abundance and the total abundance of y in the ith cell. We continue to assume units of molar concentration. Due to the stochastic nature of biochemical interactions between low concentration substrates, we expect biological variation between a sample of cell level observations $Y_1, ..., Y_n$. We model that variation with a Markov process-based description of the signaling kinetics.

To do so, we need to describe the system in terms of particles. Let c denote a constant factor that converts from units of concentration to particle counts (i.e. Avogadro's constant $\approx 6.022 * 10^{23}\text{mol}^{-1}$ times an appropriate cell volume such as 4e-12 liters). We define the function

$$\rho(u) = \lfloor cu \rfloor$$

that converts a value from continuously-valued concentrations to particle counts (integers). Then $\rho(N_{y,i})$ and $\rho(Y_i)$ are respectively the total number of particles and active particles of y in the ith cell. Suppose that due to biological stochasticity, the particle count $\rho(N_{y,i})$ varies between cells around a mean $c\mu_y$, and that in the ith

cell each of $\rho(N_{y,i})$ particles of y is enzymatically active with some probability $\theta_{y,i}$. This suggests the following data model.

$$\begin{aligned} \rho(N_{y,i}) \mid \mu_y &\sim \text{Poisson}(c\mu) \\ \rho(Y_i) \mid \rho(N_{y,i}), \theta_{y,i} &\sim \text{Binomial}(\rho(N_{y,i}), \theta_{y,i}) \end{aligned} \tag{4.27}$$

However, most experimentalists do not quantify N_y in single cell snapshot experiments, preferring to devote coverage to more proteins. Thus we assume $\rho(N_{y,i})$ is a latent variable. Marginalizing over $\rho(N_{y,i})$ in Equation 4.27, the marginal probability distribution of $\rho(Y_i)$ becomes a Poisson with mean $c\mu_y\theta_{y,i}$.

$$\rho(Y_i) \mid \mu_y, \theta_{y,i} \sim \text{Poisson}(c\mu_y\theta_{y,i}) \tag{4.28}$$

Finally, we adjust from working with particles to back to working with units of concentration. Dividing out c, we have

$$E(Y_i) = \mu_y\theta_{y,i} \tag{4.29}$$

$$\text{var}(Y_i) = \frac{\mu_y\theta_{y,i}}{c} \tag{4.30}$$

The resulting distribution is a continuous analog to a scaled-Poisson distribution with this mean and variance. The normal distribution a possible candidate distribution for modeling Y_i, since and the normal distribution is often used to approximate the Poisson. However, one problem with using a normal distribution is that the mean $\mu_y\theta_{y,i}$ is often close to 0 while Y_i cannot take negative values. Therefore we model Y_i with a gamma distribution.

$$Y_i \mid \mu_y\theta_{y,i} \sim \Gamma(c\mu_y\theta_{y,i}, c) \tag{4.31}$$

This shape $c\mu_y\theta_{y,i}$ and rate c produce the mean and variance in Equations 4.29 and 4.30.

Modeling the distribution of y based on its parents in the DAG. We defined the probability a y particle in the ith cell is active as $\theta_{y,i}$. The causal network modeling assumption is that the protein's activity is determined by the activity of

the parents in the graph. Therefore we know that $\theta_{y,i}$ should depend on the the absolute abundances of y's p parents $X_{1,i}$ ... $X_{p,i}$.

We model the system as a continuous time discrete state Markov process, then derive the value of θ_y. Assume that y has 2-state or 3-state activation as described in 4.3.2. Let $E_\pi(.)$ represent the expectation function over the distribution of $\rho(Y)$ conditional on X and $\rho(N)$. Then by *Kolmogorov's forward equation* (not shown, see [93]), the change in the expectation of $\rho(Y)$ in time is

$$\begin{aligned}\frac{d}{dt}E_\pi(\rho(Y)) &= E_\pi(vX\rho(Y^o)) - E_\pi(\alpha\rho(Y)) \\ &= vX\rho(N) - (vX+\alpha)E_\pi\rho(Y) \\ &= vX\rho(N) - (vX+\alpha)\rho(N_y)\theta_y \end{aligned} \tag{4.32}$$

Setting $\frac{d}{dt}E_\pi(\rho(Y))$ to 0 provides the steady state expectation.

$$0 = vX\rho(N_y) - (vX+\alpha)\rho(N_y)\theta_y$$

$$\theta_y = \frac{vX}{vX+\alpha} \tag{4.33}$$

$$= \frac{\beta X}{1+\beta X} = g(\beta X) \tag{4.34}$$

This identity is not true in the case of feedback loops, because of the nonlinearity introduced by the loop. However, it is true in the case of the super-node regulatory function in Equation 4.26.

4.3.4 Considerations for modeling absolute abundance

Until this point, we have assumed raw data was identical to concentration values, a quantification of absolute abundance. We derived our model by converting these values to particle counts, then multiplying by a factor c equal to Avogadro's constant times cell volume. In fact, single cell snapshot quantifications are spectra values, at best assumed to be proportional to absolute abundance, and more commonly interpreted in terms of relative abundance between samples. We also note

that there is uncertainty and variability in cell volume, though CyTOF does capture data on cell size.

We assume the quantification values are proportional to absolute abundance. We model c as a random variable. In the context of our formulation gamma distribution on Y_i, it behaves as a free parameter that can model dispersion in the data. The prior distribution on c can be informative when the experimentalists has prior knowledge on the relationship between quantifications and absolute abundance. For an uninformative prior, we use the half-Cauchy prior as suggested in [94].

Changing c results in changing the relation between the coefficients and the rate parameters. Under the mass action kinetic rate law, Equation 4.13 becomes

$$\beta = \frac{v}{c\alpha} \tag{4.35}$$

Under the Michaelis-Mention kinetic rate law, Equation 4.14 becomes

$$\beta = \frac{\tilde{v}}{c\alpha} \tag{4.36}$$

4.3.5 MCMC procedure

We propose an Bayesian MCMC sampling scheme for a given node y

$$\begin{aligned} c &\sim \text{halfCauchy}(0, 5) \\ \mu_y &\sim \pi(\mu_y) \\ \beta_{y,1}, ..., \beta_{y,p} &\sim \text{Cauchy}(0, 5) \\ \theta_{y,i} &= g(\sum_j \beta_{y,j} X_{j,i}) \\ Y_i \mid \mu_y \theta_{y,i} &\sim \Gamma(c\mu_y\theta_{y,i}, c) \end{aligned} \tag{4.37}$$

$\pi(\mu_y)$ denotes the prior on μ_y. When the experimentalist has prior knowledge of μ_y, the expected concentration of y in a cell, we set $\pi(\mu_y) = \Gamma(a, b)$ where hyperparameters a and b are specified to reflect that knowledge.

Otherwise, μ_y, along with c are simulated from the half-Cauchy distribution with hyperparameters (0, 5), the default uninformative prior for scale parameters in Stan

(see [94] for discussion). Similarly, each β_j is sampled independently from an uniformative Cauchy prior with hyperparameters (0, 5), a weakly informative prior suitable for linear coefficients [95]. Often, prior information is available in terms of the sign of β_j, i.e. positive for activation, negative for inhibition. In these cases we constrain the domain of the parameter to positive or negative real numbers in the MCMC program.

For each node y in the DAG ordering, we specify the same sampling scheme. This enables us to apply an object-oriented programing approach to building a causal Bayesian network model (as first defined in [96]). Each node is an instance from a class where attributes include a given target node, its parent nodes in the network, a regulatory function, and optionally, an informative prior distribution on each prior. The overall sampler is compiled by iterating through nodes according to the topological ordering for the DAG.

4.4 Data

4.4.1 ODE Model of MAPK signaling

We demonstrate the Bayesian network modeling approach using the Huang-Ferrell model of the MAPK signaling cascade [26]. In the model E1 represents is a fixed value representing signal in the system. E1 is an enzyme that catalyzes the phosphorylation of Raf, phosphorylated Raf catalyzes the double-phosphorylation of Mek, and doubly-phosphorylated Mek catalyzes the double phosphorylation of Erk. Doubly-phosphorylated Erk is a key signaling kinase that regulates transcription responses such as proliferation. The model illustrates the "ultrasensitive" behavior of the pathway, i.e. that Erk activation is strong even at low E1 signals.

The model represents the phosphorylation and dephosphorylation reactions of each species in the cascade with Michaelis-Menton kinetics. Each phosphorylation and dephosphorylation reaction has the same basic set of parameters, shown in Table 4.1. For simplicity, the parameter values are the same for each reaction.

Table 4.1.
Reaction parameters in MAPK model

Reaction Type	Parameters	Values
Binding	k_{bind}	1000
Unbinding	k_{unbind}	150
Transformation	v	150
MM-Constant	$K = \frac{v+k_{\text{unbind}}}{k_{\text{bind}}}$	0.3

We used the model to simulate an experiment with 5 samples of 100,000 single cell observations, each with different levels of initial E1 concentration. We used a stochastic simulation algorithm to simulate cell trajectories [97]. Each snapshot measurement is a steady-state measurement of the system (after 100 seconds).

4.4.2 Mass cytometry data from study of T-cell signaling

We worked with the dataset using single cell level quantifications of intracellular signaling proteins in CD4+ naive T-cells. The signaling mechanism within immune cells is a finely-tuned classifier of environmental signal. It must be sensitive enough to detect rare signals from foreign antigens, yet specific enough not to misclassify as threats the vast majority of incoming signals comes from "self"-related species (as in autoimmune diseases). Upon processing the signal they must decide between one of several possible downstream signaling responses, from proliferation, differentiation, to senescence.

Several of the signaling proteins involved in this signal processing mechanism were quantified in a mass cytometry study of T-lymphocyte populations in B6 mice [86]. In the study TCR is stimulated with two different types of stimuli (CD3/CD28 and CD3/CD4/CD28), and samples are collected at 13 time points after TCR activation ranging from 30 s to 80 min. 10,000 cells were quantified within each sample. Surface markers were used to distinguish cells into six T cell subsets.

We focus on naïve CD4+ cells, and examine the functional relationship between CD3z, and SLP76. CD3z together with TCR activates downstream pathways upon antigen activation. When phosphorylated CD3z (upon recruitment of ZAP-70) SLP76 (src homology 2 domain-containing leukocyte phosphoprotein). Loss of SLP-76 results in a near total loss of TCR signal transduction [98]. We examine cells measured at the 30 second time point, a point at which the decision mechanism reaches a quasi-steady state. We fit our model and compare our model fit to signal-response curve fitting approaches that, unlike our model, do not provide insight

Table 4.2.
Coefficient parameters in MAPK model

Relation	Phosphotase	Activation	Deactivation	coefficient
E1 $\rightarrow$ Raf	$C_{\text{Raf}} = 722$	$v_1 = \frac{v}{K}$	$\alpha_1 = \frac{vC_{\text{Raf}}}{K}$	$\beta_1 = \frac{v_1}{\alpha_1}$
Raf $\rightarrow$ Mek	$C_{\text{Mek}} = 722$	$v_2 = \frac{v}{K}$	$\alpha_2 = \frac{vC_{\text{Mek}}}{K}$	$\beta_2 = \frac{v_2}{\alpha_2}$
Mek $\rightarrow$ Erk	$C_{\text{Erk}} = 289062$	$v_3 = \frac{v}{K}$	$\alpha_3 = \frac{vC_{\text{Erk}}}{K}$	$\beta_3 = \frac{v_3}{\alpha_3}$

into underlying kinetic rates or quantify uncertainty on the estimates of the curve function's parameters.

4.4.3 Code and materials

The analysis and workflows described in this work were implemented in R. We conducted the Bayesian modeling and inference using the Stan probabilistic programming language [99]. All workflows and analysis code described in this work are available as an R package available online at *github.com/robertness/signalnet*.
The Huang-Ferrell model used in this package was sourced from the online BioModel's database [100], then modified to reflect first order Michaelis-Menton kinetics. Synthetic datasets were simulated from the modified model using the COPASI pathway simulation tool [101]. The modified model files and synthetic datasets are included in the R package. The CyTOF data from the T cell study was sourced from the investigators' website. R workflows for loading, processing, and conducting the statistical analysis on the data are included in the R package.

4.5 Results

4.5.1 The inference procedure recovered stable posterior densities centered around the true values of the Huang-Ferrell MAPK model

We applied our inference procedure to the synthetic single cell snapshot data simulated from the Huang-Ferrell model. We target the coefficients parameters β_1 (E1 regulation of Raf), β_2 (Raf regulation of Mek), and β_3 (Mek regulation of Erk). Each coefficient is a ratio of an activation rate parameter v_i and a deactivation rate parameter α_i, which themselves are functions of the reaction parameters in Table 4.1. The model maintains simplicity by assigning the same values to the underlying parameters for each reaction, such that the coefficients simplify in this case to the inverse of phosphotase abundance. Table 4.2 shows the relationship between reaction parameters and coefficients.

We implement the MCMC scheme in the probabilistic programming language Stan [92] We use the conditional expectation function for single phosphorylation (Equation 4.16) for E1 activation of Raf and double phosphorylation (Equation 4.19) for Raf activation of Mek and Mek activation of Erk. We assume the conservation rules are not known, but are constant across cells. The full model was as follows

$$\begin{aligned}
\mu_{Raf}, \mu_{Mek}, \mu_{Erk} &\sim \text{halfCauchy}(0, 5) \\
\beta_1, \beta_2, \beta_3 &\sim \text{Cauchy}(0, 5) \\
c &= n_a * V \\
\text{Raf}_i &\sim \Gamma(c\mu_{Raf} g_1(\beta_1 \text{E1}_i), c) \\
\text{Mek}_i &\sim \Gamma(c\mu_{Mek} g_2(\beta_2 \text{Raf}), c) \\
\text{Erk}_i &\sim \Gamma(c\mu_{Erk} g_2(\beta_3 \text{Mek}), c)
\end{aligned} \tag{4.38}$$

where n_a is Avogadro's constant and V is the cell volume specified in the model. Figure 4.2 illustrates the posterior densities of the coefficient parameters, which centered symmetrically around the true values of the parameter (solid line). In each

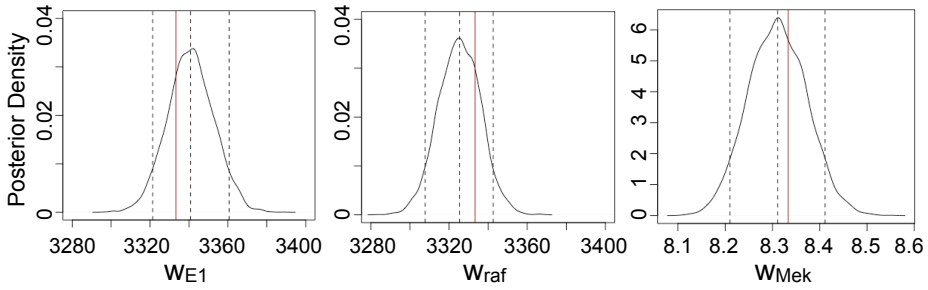

Fig. 4.2. Density of simulated values from posterior of the coefficient densities. Dashed lines show the mean and the .05, .95 percentiles. Solid lines show the true value of the coefficient parameter. The posterior densities of the coefficient are stable and centered around the true values.

case the densities are stable and enabled basic inference; the dashed lines illustrate the 10% credible interval for the model estimates.

4.5.2 Coefficients can inform time course experiments

We demonstrate the use of single cell snap shot data as a means of improving direct inference of rate parameters from time course experiments. In the Huang-Ferrell MAPK model, the phosphorylation of kinase Raf is catalyzed by an input stimulus E1. The regulation of Raf depends on the rate parameters for activation (v_1) and deactivation (α_1), listed in Table 4.2. Under first-order Michaelis-Menton kinetics

$$v_1 = \frac{v}{K} = \frac{v}{\frac{v+k_{\text{unbind}}}{k_{\text{bind}}}} = 500 \tag{4.39}$$

$$\alpha_1 = [\text{Raf-tase}] * v = 0.15 \tag{4.40}$$

where K is the Michaelis-Menton constant and [Raf-tase] is concentration of Raf phosphatase, which in this model is assumed constant at 0.0003 μmol/L.
Inference of v_1 and α_1 from data requires temporal quantifications of phosphorylated Raf (p-Raf) concentrations after the introduction of stimulus, such as from a fluorescence microscopy study. These parameters are not identifiable from snapshot data, such as with CyTOF. However, $\beta_1 = \frac{v_1}{v_2}$ is identifiable with snapshot data. We demonstrate with experiments simulated from this model how inference of β_1 from snapshot data can inform the inference of v_1 and α_1 in a follow up experiment that collects time course measurements. Supposed that in a time course experiment, we assign log-normal priors to v_1, α_1, and use the following sampling scheme in MCMC;

$$\begin{aligned}
\mu_\alpha, \mu_v &\sim \text{Cauchy}(0, 5) \\
\sigma_\alpha, \sigma_v &\sim \text{half-Cauchy}(0, 5) \\
\alpha_1 &\sim lnN(\mu_\alpha, \sigma_\alpha^2) \\
v_1 &\sim lnN(\mu_v, \sigma_v^2)
\end{aligned} \tag{4.41}$$

where μ_v, σ_v, μ_α, and σ_α are hyperparameters that capture prior knowledge on the location and spread of the rates. We further assume that β_1 is also lognormal and independent of α_1 (true if v_1 is unknown). By the properties of the lognormal distribution, we have the following;

$$\begin{aligned} \mu_\alpha &\sim \text{Cauchy}(0, 5) \\ \sigma_\alpha &\sim \text{half-Cauchy}(0, 5) \\ \alpha_1 &\sim lnN(\mu_\alpha, \sigma_\alpha^2) \\ v_1 \mid \alpha_1, \beta_1 &\sim lnN(\mu_\alpha + \mu_\beta, \sigma_\alpha^2 + \sigma_\beta^2) \end{aligned} \tag{4.42}$$

where μ_β, and σ_β are hyperparameters for the prior on β. Sampling statement 4.42 replaces 4.41, eliminating hyperparameters μ_v, σ_v, since the location and scale of v_1 are now determined by μ_α, σ_α, μ_β, and σ_β. We infer μ_β, σ_β from the steady state data, i.e. using information in the steady-state data to construct an informative prior on v_1.

We simulated a experiment with 1000 snapshots for each of 10 input stimulus concentrations ranging across measurements, for 10000 snapshots of steady state signaling, which is a conservative number of cells in a single cell proteomics experiment. Figure 4.3 A illustrates phosphorylated Raf (p-Raf) concentrations for each level of input signal such that variation increases in as p-Raf concentration increases. We simulated a second experiment capturing 3 time courses (Figure 4.3 B), one for each of the same initial stimulus. Each time course lasts 60 seconds, time points collected every 5 seconds.

Figure 4.3 C and D illustrate two cases of posterior inference of v_1 and v_2. The light grey posterior densities illustrate stand-alone inference of the parameters from the time course data. The dark grey posterior distributions illustrate posterior inference when we first do inference on w from the snapshot data, and use those results to constrain the inference of v_1 and v_2 with the time course data. The latter case produces posterior densities with less variation.

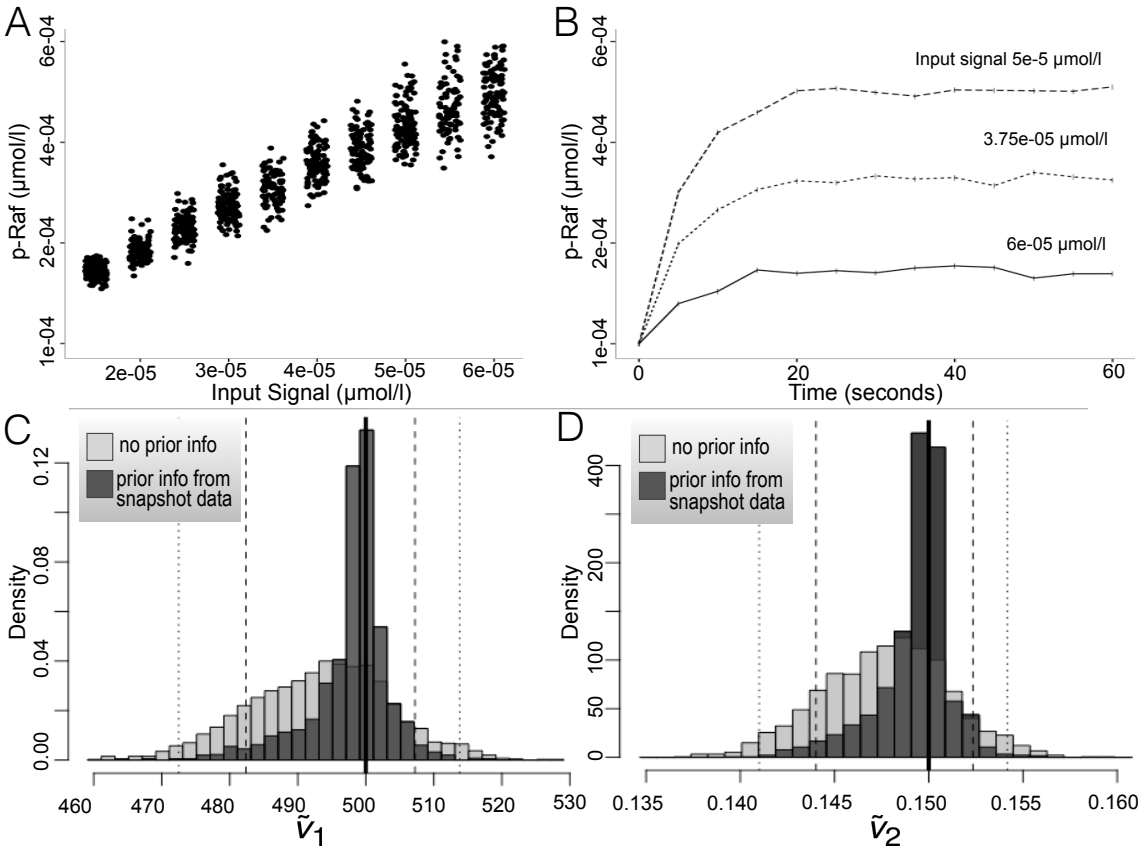

Fig. 4.3. Inference of rate parameters for Raf kinase activation (v_1) and deactivation (α_1) from simulated experiments from the Huang Ferrell model of MAPK signaling. A: Raf concentration given increasing levels of input signal. B: 3 sample time courses of Raf given increasing input signal. C and D: Comparison of parameter posterior distributions with and without priors informed by inference on the snapshot data. The informative prior improves precision in the posterior.

This preliminary result motivates the use of the proposed approach as a tool for experimental design. Single cell experiments have the throughput to quantify multiple nodes simultaneously, enabling simultaneous inference of multiple coefficient parameters. This produces a causal Bayesian network model capable of predicting the steady state of any set of proteins given the states of others in the network. Time course experiments have lower throughput, enabling only a small subset of the βs to be parsed into rate parameters. The posteriors of the coefficient parameters given single cell experiments can motivate various design strategies for a follow up time course experiments depending on the utility criterion the investigator seeks to optimize. For example, if inference on specific set of rate parameters is of interest, the results of the inference with the single cell experiment can be used to increase precision in the results of the parameter inference with the time course experiment. Alternatively, suppose one intends to use the causal network for a specific prediction task, and the results inference of the coefficient parameters on single cell data reveal that he prediction are highly sensitive to the values of a certain coefficient parameters. One could design a follow up time course experiment that improves the precision of the important rates distribution by inferring the rate parameters that comprise them.

4.5.3 The inference procedure performs as well as nonparametric response curve modeling on CyTOF data

We fit the proposed Bayesian model. Figure 4.4 shows the results of the the proposed model fit on the single cell measurements in the T-cell signaling data. The figure shows a conditional density heat map of SLP76 quantifications given CD3z quantifications, generated using conditional kernel density estimation similar to the approach described in Krishnaswamy et al. [86]. Across the range of CD3z quantifications, we calculated the posterior mode of SLP76, and superimposed the resulting curve over the heat map. The curve coincides with the dense regions of the map,

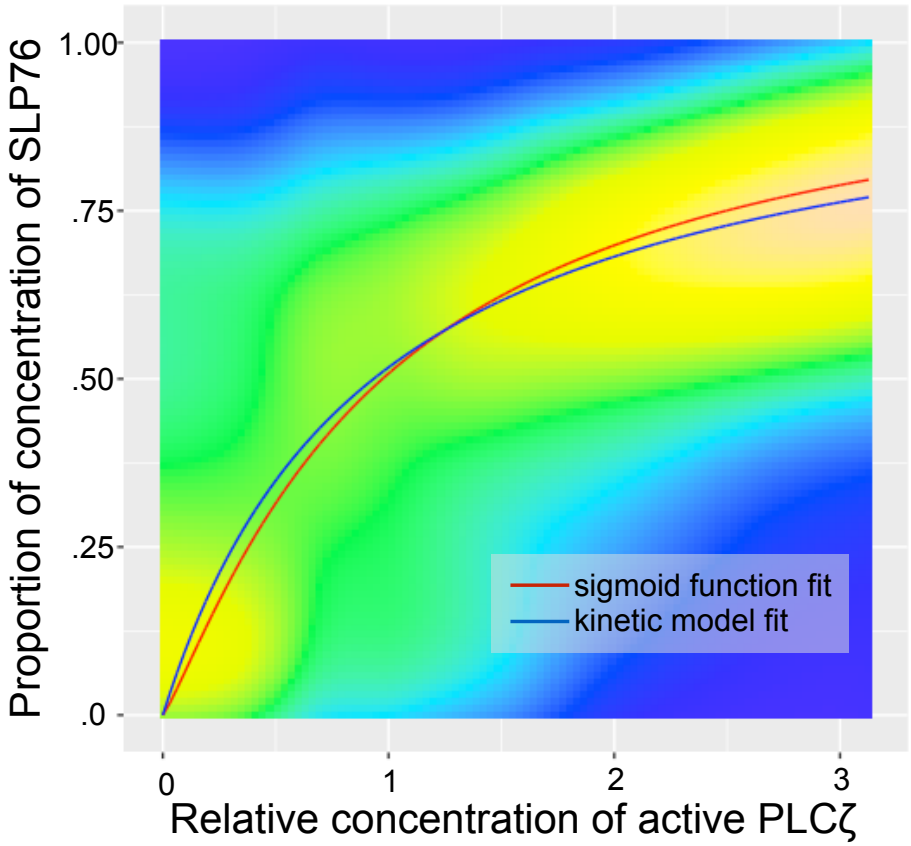

Fig. 4.4. The visualization of the conditional density of SLP76 given CD3ζ fit using conditional kernel density estimation. The lines represent estimation of the signal-response function by curve-fitting the data. The red line is a general sigmoidal function fit with nonlinear regression. The blue line represents the proposed kinetic model. In addition to having a comparable fit, the model's parameter estimates are interpretable in terms of rate parameters.

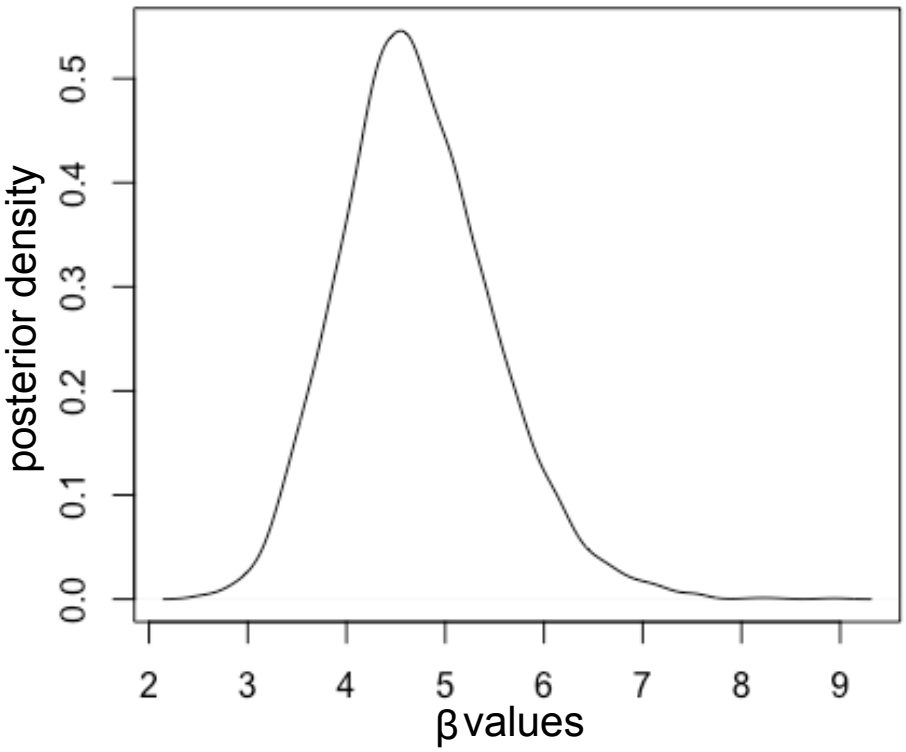

Fig. 4.5. Posterior densiity of the coefficient estimate. The posterior mean is 4.72. The 95% credible interval (highest posterior density interval) is $\{3.25, 6.22\}$. Assuming the mass action rate law, on average the rate parameter for activation is 4.72 times that of c * the rate parameter for deactivation. With more information on c, we could make more direct assessments on the ratio of rate parameters.

and is comparable to a sigmoid function fit to the modes of the conditional kernel density estimate. The results illustrate the Bayesian approach attains a comparable signal-response fit.

Figure 4.5 illustrates the posterior density of the coefficient estimate. The posterior mean is 4.72, and the 95% credible interval (highest posterior density interval) is $\{3.25, 6.22\}$. I we were to assume the mass action rate law and that c were known, the average the rate parameter for activation is 4.72 times that of the rate parameter for deactivation. Otherwise it is $4.72/c$ times more. With more information on c, we could make more direct assessments on the ratio of rate parameters. The

posterior probability of any hypotheses concerning the coefficient parameter can be evaluated against the posterior, and highly probable and interesting hypotheses can be assessed in targeted validation experiments.

4.6 Discussion

Our results showed that we can characterize the dynamics of cell signaling from multiple single cell instantaneous proteomic "snapshots" of the system. Using this approach we can fit a model of cell signaling that quantifies causal influence in terms of the underlying rates of the biochemical reactions underlying signaling, as well as perform traditional graphical modeling functions such as prediction and probabilistic inference.

Our proposed approach assumes mass action or Michaelis-Menton kinetics, though we see no reason why more complex kinetic assumptions could not be used. The key assumption is that the signaling system has reached a steady-state, or at least a stable quasi-steady state that is preserved for a duration of time relevant to the biological question. Though this makes the approach inappropriate for some signaling mechanisms (eg. oscillators), our results demonstrate the approach works with signaling pathways relevant to important therapeutic domains such as immunology and oncology.

Our approach lays the groundwork for biological interpretation from coefficient parameters. A key assumption of our model is that the quantification measurements in single cell proteomic technologies such as CyTOF are proportional to the absolute abundance of a given protein in the cell, and that the mapping from quantifications values to absolute abundance is known. In standard practice for both flow cytometry and CyTOF, raw single cell data is normalized (eg based on a bead standard for CyTOF [102]) then subjected to a scale transformation such as the inverse hyperbolic function [103, 104]. Transformation methods facilitate visualization and analysis of the data, as well as reduce noise in measurements close to the limits of

detection [103, 105]. To our knowledge, there has been little examination of the relationship between these procedures and the information about absolute abundance in the transformed data, for example if it makes it difficult to detect key nonlinear relationships between proteins. We believe the modeling approach outlined in this work provides motivation for work in this area.

In absence of a mapping of quantifications to absolute abundance, the proposed approach still provides a causal Bayesian network where the causal influence a node receives from its parents in the network is quantified in terms of a coefficient parameter. This approach compares favorable to a nonparametric approaches proposed in prior literature [86]. Further, it allows for simultaneous fitting of an entire network, rather than fitting each parent-child relationship separately.

We demonstrated the potential for our approach to inform low-throughput experiments where as small set of proteins are quantified from a single cell over several time points in during the transient pre-steady state state of signaling. Another alternative would be to collect single snapshot measurements at multiple time points after stimuli. This requires selection of a set of time points for fixing cells (freezing the state of signaling in a sample prior to submitting it to the measurement device), then fixing those cells in each replicate for each combination of conditions. Selecting time points that are too few or not close enough enough in time would not be sufficient to resolve transient dynamics, but the complexity of the experiment increases in the number of selected time points. Our proposed approach could provide information about the transience dynamics of the system that could be used to inform the selection of time points.

Our results demonstrated the potential for an iterative model building workflow for proteomics. In the first step one uses high-throughput mass spectrometry techniques to identify key proteins in the signaling phenotype. In the second step, one uses network inference techniques with targeted, high-sampling throughput experiments or single cell proteomics experiments to learn directed network structure. In the fourth step, one applies the proposed approach to inferring coefficient param-

eters from single cell data, where coefficients are functions of biochemical rate parameters. Finally, the results from inference on the rates are used to inform follow up time course experiments that target the rate parameters specifically. We note that single-cell experiments do not eliminate the need for true biological replication. The proposed workflow can be extended to population level inference by modeling subjects as additional nodes in the network.

Overall, we believe that the proposed strategy is an important step towards practical inference of causal networks, and advocate its practical use.

5. SUMMARY AND FUTURE WORK

This dissertation focuses on the problem of causal inference of cell signal transduction. The objective is to infer a causal Bayesian network model of signal transduction from proteomics data. The dissertation contributes methods that support the three causal inference tasks in pursuit of this objective: (1) inferring the presence of causal edges; (2) inferring the direction of the causal edge conditional on its presence; and (3) inferring the magnitude of causal influence conditional on the presence and direction of the edge.

This dissertation focused on the experimental settings needed for each of these tasks, as well as Bayesian methods for conducting the inference. As an experimental platform, we focused on snapshot data, i.e. data where each sample quantifies signal transduction in the system at a single time point. We payed special attention to single cell experiments, where each quantification is a cell-level replicate of cell signaling, and contrasted it with bulk (non-single cell) experiments.

Contributions to task 1. Machine learning algorithms for detecting the presence of causal relationships rely on algorithms for detecting conditional independence. This dissertation includes a simulation analysis of the relationship between the dimensionality of experimental data and the detection of edges when applying this algorithm. The results demonstrated that the best setting for edge detection were experiments that focused on a limited set of proteins where one has prior knowledge of signaling relationships. The dissertation provides guidelines for designing such discovery experiments.

Contributions to task 2. This dissertation proposes an active learning strategy for selecting targeted interventions that resolve causal edge orientation. The method takes as an input prior knowledge on edge presence and orientation from

pathway databases such as KEGG. It combines this prior knowledge with historic datasets to create a prior probability distribution on the space of graphs. The method uses this distribution to prioritize targeted interventions by the expected number of edges that will be oriented by their application in a causal inference experiment.

Contributions to task 3. This dissertation casts causal influence as a problem of estimating the parameters of the conditional probability distribution in a causal Bayesian network. The magnitude of the parameter quantifies the magnitude of causal influence. Moreover, we constructed the probability distribution based on the underlying statistical mechanics of signaling reactions, such that the entropy in the conditional probability distribution reflects the biological stochasticity that is evident in the variation across cells. The estimated parameters are estimable functions of the rates. We show how the posterior distribution on these parameters can improve direct inference of the rate parameters from low-throughput time course data.

Overall, this dissertation provides a workflow for building a causal Bayesian network model of signaling from high-throughput proteomics experiments. In future work, one experimental and modeling workflow can unite these three tasks. Starting from high-throughput experiments that discover components of the pathway, to experiments that target fewer proteins, and apply more perturbations (both general and targeted interventions) that aid in the detection of the presence, orientation, and kinetics of causal regulatory relationships. As a sequential Bayesian method, each step can directly incorporate the results of the previous, rather than having to construct prior distributions from scratch at each step.

REFERENCES

REFERENCES

[1] H. Lodish, D. Baltimore, A. Berk, S. L. Zipursky, P. Matsudaira, and J. Darnell, *Molecular cell biology*. Scientific American Books New York, 1995, vol. 3.

[2] C. Terfve and J. Saez-Rodriguez, "Modeling signaling networks using high-throughput phospho-proteomics," in *Advances In Systems Biology*. Springer, 2012, pp. 19–57.

[3] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, no. 1, p. 565, 2012.

[4] D. Eaton and K. P. Murphy, "Exact Bayesian structure learning from uncertain interventions," in *International Conference On Artificial Intelligence and Statistics*, 2007, pp. 107–114.

[5] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[6] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge Univ Press, 2000, vol. 29.

[7] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, 2002.

[8] N. Friedman, "Inferring cellular networks using probabilistic graphical models," *Science*, vol. 303, no. 5659, pp. 799–805, 2004.

[9] F. Markowetz and R. Spang, "Inferring cellular networks–a review," *BMC bioinformatics*, vol. 8, no. Suppl 6, p. S5, 2007.

[10] J. Pearl, *Causality*. Cambridge University Press, 2009.

[11] 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 (New York, N.Y.)*, vol. 308, no. 5721, pp. 523–9, Apr. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15845847

[12] F. Eberhardt, C. Glymour, and R. Scheines, "On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables," *arXiv Preprint arXiv:1207.1389*, 2012.

[13] Y.-B. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[14] S. Meganck, P. Leray, and B. Manderick, "Learning causal Bayesian networks from observations and experiments: A decision theoretic approach," in *Modeling Decisions for Artificial Intelligence*. Springer, 2006, pp. 58–69.

[15] K. P. Murphy, "Active learning of causal Bayes net structure," 2001.

[16] I. Pournara and L. Wernisch, "Reconstruction of gene networks using Bayesian learning and manipulation experiments," *Bioinformatics*, vol. 20, no. 17, pp. 2934–2942, 2004.

[17] J. Tian and J. Pearl, "Causal discovery from changes," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 512–521.

[18] S. Tong and D. Koller, "Active learning for structure in Bayesian networks," in *International Joint Conference On Artificial Intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 863–869.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the Graphical LASSO," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[20] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation," *Nucleic Acids Researchearch*, vol. 44, no. D1, pp. D457–D462, 2016.

[21] K. Sachs, S. Itani, J. Fitzgerald, B. Schoeberl, G. Nolan, and C. Tomlin, "Single timepoint models of dynamic systems," *Interface focus*, vol. 3, no. 4, p. 20130019, 2013.

[22] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt, and P. K. Sorger, "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction," *Molecular systems biology*, vol. 5, no. 1, p. 331, 2009.

[23] R. J. Prill, J. Saez-Rodriguez, L. G. Alexopoulos, P. K. Sorger, and G. Stolovitzky, "Crowdsourcing network inference: the dream predictive signaling network challenge," *Science signaling*, vol. 4, no. 189, p. mr7, 2011.

[24] A. Bensimon, A. J. Heck, and R. Aebersold, "Mass spectrometry-based proteomics and network biology," *Annual review of biochemistry*, vol. 81, pp. 379–405, 2012.

[25] T. Holbro and N. E. Hynes, "Erbb receptors: directing key signaling networks throughout life," *Annu. Rev. Pharmacol. Toxicol.*, vol. 44, pp. 195–217, 2004.

[26] C.-Y. Huang and J. E. Ferrell, "Ultrasensitivity in the mitogen-activated protein kinase cascade," *Proceedings of the National Academy of Sciences*, vol. 93, no. 19, pp. 10 078–10 083, 1996.

[27] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000, vol. 81.

[28] M. Scutari, "Learning bayesian networks with the bnlearn r package," *arXiv preprint arXiv:0908.3817*, 2009.

[29] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, no. 2, pp. 293–314, 2014.

[30] D. Margaritis, "Learning bayesian network model structure from data," Ph.D. dissertation, US Army, 2003.

[31] I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS Conference*, vol. 2003, 2003, pp. 376–381.

[32] S. Yaramakala and D. Margaritis, "Speculative markov blanket discovery for optimal feature selection," in *Data mining, fifth IEEE international conference on.* IEEE, 2005, pp. 4–pp.

[33] A. Hauser and P. Bühlmann, "Two optimal strategies for active learning of causal models from interventional data," *arXiv preprint arXiv:1205.4174*, 2012.

[34] A. Hyttinen, F. Eberhardt, and P. O. Hoyer, "Experiment selection for causal discovery," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3041–3071, 2013.

[35] J. Saez-Rodriguez and C. D. A. Terfve, *Modeling Signaling Networks Using High-throughput Phospho-proteomics* , ser. Advances in Experimental Medicine and Biology, I. I. Goryanin and A. B. Goryachev, Eds. New York, NY: Springer New York, 2012, vol. 736.

[36] L. C. Gillet, P. Navarro, S. Tate, H. Roest, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis," *Molecular & Cellular Proteomics*, 2012.

[37] 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*, vol. 308, no. 5721, pp. 523–529, 2005.

[38] J. Saez-Rodriguez, L. G. Alexopoulos, J. Epperlein, R. Samaga, D. A. Lauffenburger, S. Klamt, and P. K. Sorger, "Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction," *Molecular Systems Biology*, vol. 5, no. 1, p. 331, Dec. 2009.

[39] C. D. A. Terfve, E. H. Wilkes, P. Casado, P. R. Cutillas, and J. Saez-Rodriguez, "Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data," *Nature Communications*, vol. 6, pp. 1–11, Sep. 2015.

[40] L. G. Alexopoulos, J. Saez-Rodriguez, B. D. Cosgrove, D. A. Lauffenburger, and P. K. Sorger, "Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes," *Molecular & Cellular Proteomics*, vol. 9, no. 9, pp. 1849–1865, 2010.

[41] T. Pawson and N. Warner, "Oncogenic re-wiring of cellular signaling pathways," *Oncogene*, vol. 26, no. 9, pp. 1268–1275, 2007.

[42] O. D. Perez and G. P. Nolan, "Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry," *Nature Biotechnology*, vol. 20, no. 2, pp. 155–162, 2002.

[43] D. R. Bandura, V. I. Baranov, O. I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J. E. Dick, and S. D. Tanner, "Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry," *Analytical Chemistry*, vol. 81, no. 16, pp. 6813–6822, 2009.

[44] T. J. Chen and N. Kotecha, "Cytobank: Providing an analytics platform for community cytometry data analysis and collaboration," in *High-Dimensional Single Cell Analysis*. Springer, 2014, pp. 127–157.

[45] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical Applications In Genetics and Molecular Biology*, vol. 6, no. 1, 2007.

[46] 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*, vol. 15, no. 3, pp. 683–690, 2016.

[47] K. B. Korb and a. E. Nicholson, *Bayesian Artificial Intelligence*. CRC Press, 2010.

[48] S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards, *Artificial Intelligence: A Modern Approach*. Prentice Hall Upper Saddle River, 2003, vol. 2.

[49] M. Scutari, "On the prior and posterior distributions used in graphical modelling," *Bayesian Analysis*, vol. 8, no. 3, pp. 505–532, 2013.

[50] G. F. Cooper and C. Yoo, "Causal discovery from a mixture of experimental and observational data," in *Proceedings of the Fifteenth Conference On Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 116–125.

[51] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, 1995.

[52] N. Friedman, M. Goldszmidt, and A. Wyner, "Data analysis with Bayesian networks: A bootstrap approach," in *Proceedings of the Fifteenth Conference On Uncertainty In Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 196–205.

[53] S. Imoto, S. Y. Kim, H. Shimodaira, S. aburatani, K. Tashiro, S. Kuhara, and S. Miyano, "Bootstrap analysis of gene networks based on Bayesian networks and nonparametric regression," *Genome Informatics*, vol. 13, pp. 369–370, 2002.

[54] N. Friedman and D. Koller, "Being Bayesian about network structure A Bayesian approach to structure discovery in Bayesian networks," *Machine Learning*, vol. 50, no. 1-2, pp. 95–125, 2003.

[55] D. M. Chickering, "A transformational characterization of equivalent Bayesian network structures," in *Proceedings of the Eleventh Conference On Uncertainty In Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 87–98.

[56] D. Rossell and P. Müller, "Sequential stopping for high-throughput experiments," *Biostatistics*, vol. 14, no. 1, pp. 75–86, 2013.

[57] Y. Guan, M. Dunham, a. Caudy, and O. Troyanskaya, "Systematic planning of genome-scale experiments in poorly studied species," *PLoS Comput Biol*, vol. 6, no. 3, p. e1000698, 2010.

[58] R. D. King, K. E. Whelan, F. M. Jones, P. G. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.

[59] R. Castelo and A. Siebes, "Priors on network structures. Biasing the search for Bayesian networks," *International Journal of approximate Reasoning*, vol. 24, no. 1, pp. 39–57, 2000.

[60] J. S. Ide and F. G. Cozman, "Random generation of Bayesian networks," in *Advances In Artificial Intelligence*. Springer, 2002, pp. 366–376.

[61] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013.

[62] D. M. Chickering and D. Heckerman, "Efficient approximations for the marginal likelihood of bayesian networks with hidden variables," *Machine learning*, vol. 29, no. 2-3, pp. 181–212, 1997.

[63] N. Friedman *et al.*, "Learning belief networks in the presence of missing values and hidden variables," in *ICML*, vol. 97, 1997, pp. 125–133.

[64] C. Terfve, T. Cokelaer, D. Henriques, A. MacNamara, E. Goncalves, M. K. Morris, M. van Iersel, D. A. Lauffenburger, and J. Saez-Rodriguez, "Cellnoptr: A flexible toolkit to train protein signaling networks to data using multiple logic formalisms," *BMC systems biology*, vol. 6, no. 1, p. 1, 2012.

[65] K. Sachs, A. J. Gentles, R. Youland, S. Itani, J. Irish, G. P. Nolan, and S. K. Plevritis, "Characterization of patient specific signaling via augmentation of bayesian networks with disease and patient state nodes," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 6624–6627.

[66] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden *et al.*, "The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

[67] N. Le Novere, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro *et al.*, "Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D689–D691, 2006.

[68] F. J. Bruggeman and H. V. Westerhoff, "The nature of systems biology," *TRENDS in Microbiology*, vol. 15, no. 1, pp. 45–50, 2007.

[69] J. J. Tyson, K. C. Chen, and B. Novak, "Sniffers, buzzers, toggles and blinkers: Dynamics of regulatory and signaling pathways in the cell," *Current Opinion in Cell Biology*, vol. 15, no. 2, pp. 221–231, 2003.

[70] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC press, 2006.

[71] N. J. Brunel *et al.*, "Parameter estimation of ode's via nonparametric estimators," *Electronic Journal of Statistics*, vol. 2, pp. 1242–1267, 2008.

[72] D. J. Wilkinson, "Parameter inference for stochastic kinetic models of bacterial gene regulation: A bayesian approach to systems biology," in *Proceedings of 9th Valencia International Meeting on Bayesian Statistics. Oxford University Press*, 2010, pp. 679–705.

[73] ——, "Stochastic modelling for quantitative description of heterogeneous biological systems," *Nature Reviews Genetics*, vol. 10, no. 2, pp. 122–133, 2009.

[74] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods," *Genome research*, vol. 13, no. 11, pp. 2467–2474, 2003.

[75] M. Quach, N. Brunel, and F. d'Alché Buc, "Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference," *Bioinformatics*, vol. 23, no. 23, pp. 3209–3216, 2007.

[76] M. D. Brennan, R. Cheong, and A. Levchenko, "How information theory handles cell signaling and uncertainty," *Science*, vol. 338, no. 6105, pp. 334–335, 2012.

[77] A. Levchenko and I. Nemenman, "Cellular noise and information transmission," *Current opinion in biotechnology*, vol. 28, pp. 156–164, 2014.

[78] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, "Stochastic gene expression in a single cell," *Science*, vol. 297, no. 5584, pp. 1183–1186, 2002.

[79] P. J. Goss and J. Peccoud, "Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets," *Proceedings of the National Academy of Sciences*, vol. 95, no. 12, pp. 6750–6755, 1998.

[80] D. T. Gillespie, *Markov processes: an introduction for physical scientists*. Elsevier, 1991.

[81] E. J. Bennett, J. Rush, S. P. Gygi, and J. W. Harper, "Dynamics of cullin-ring ubiquitin ligase network revealed by systematic quantitative proteomics," *Cell*, vol. 143, no. 6, pp. 951–965, 2010.

[82] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling, "Ensemble modeling for analysis of cell signaling dynamics," *Nature biotechnology*, vol. 25, no. 9, pp. 1001–1006, 2007.

[83] C. Ludwig, M. Claassen, A. Schmidt, and R. Aebersold, "Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry," *Molecular & Cellular Proteomics*, vol. 11, no. 3, pp. M111–013 987, 2012.

[84] A. C. Pfeifer, D. Kaschek, J. Bachmann, U. Klingmüller, and J. Timmer, "Model-based extension of high-throughput to high-content data," *BMC systems biology*, vol. 4, no. 1, p. 106, 2010.

[85] K. Sachs, D. Gifford, T. Jaakkola, P. Sorger, and D. A. Lauffenburger, "Bayesian network approach to cell signaling pathway modeling," *Sci STKE*, vol. 148, p. e38, 2002.

[86] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau, S. C. Bendall, O. Litvin, E. Stone, D. Pe'er, and G. P. Nolan, "Conditional density-based analysis of t cell signaling in single-cell data," *Science*, vol. 346, no. 6213, p. 1250689, 2014.

[87] J. Selimkhanov, B. Taylor, J. Yao, A. Pilko, J. Albeck, A. Hoffmann, L. Tsimring, and R. Wollman, "Accurate information transmission through dynamic biochemical signaling networks," *Science*, vol. 346, no. 6215, pp. 1370–1373, 2014.

[88] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[89] D. J. Wilkinson, "Bayesian methods in bioinformatics and computational systems biology," *Briefings in bioinformatics*, vol. 8, no. 2, pp. 109–116, 2007.

[90] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

[91] M. Plummer *et al.*, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd international workshop on distributed statistical computing*, vol. 124. Vienna, 2003, p. 125.

[92] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *J Stat Softw*, 2016.

[93] N. G. Van Kampen, *Stochastic processes in physics and chemistry*. Elsevier, 1992, vol. 1.

[94] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian Analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[95] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, pp. 1360–1383, 2008.

[96] D. Koller and A. Pfeffer, "Object-oriented bayesian networks," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1997, pp. 302–313.

[97] M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," *The journal of physical chemistry A*, vol. 104, no. 9, pp. 1876–1889, 2000.

[98] J. E. Smith-Garvin, G. A. Koretzky, and M. S. Jordan, "T-cell activation," *Annual Review of Immunology*, vol. 27, p. 591, 2009.

[99] A. Gelman, "Rstan: the r interface to stan," 2014.

[100] N. Juty, R. Ali, M. Glont, S. Keating, N. Rodriguez, M. Swat, S. Wimalaratne, H. Hermjakob, N. Le Novre, C. Laibe *et al.*, "Biomodels: Content, features, functionality, and use," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 4, no. 2, pp. 55–68, 2015.

[101] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer, "Copasi—a complex pathway simulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[102] R. Finck, E. F. Simonds, A. Jager, S. Krishnaswamy, K. Sachs, W. Fantl, D. Pe'er, G. P. Nolan, and S. C. Bendall, "Normalization of mass cytometry data with bead standards," *Cytometry Part A*, vol. 83, no. 5, pp. 483–494, 2013.

[103] S. C. Bendall, E. F. Simonds, P. Qiu, D. A. El-ad, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky *et al.*, "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum," *Science*, vol. 332, no. 6030, pp. 687–696, 2011.

[104] D. R. Parks, M. Roederer, and W. A. Moore, "A new "logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data," *Cytometry Part A*, vol. 69, no. 6, pp. 541–551, 2006.

[105] K. O'Neill, N. Aghaeepour, J. Špidlen, and R. Brinkman, "Flow cytometry bioinformatics," *PLoS Comput Biol*, vol. 9, no. 12, p. e1003365, 2013.

VITA

VITA

Robert Donald Osazuwa Ness III

Department of Statistics, Purdue University

150 N. University St. West Lafayette IN 47907 U.S.A.

email: nessr@purdue.edu

Education

2016	PH.D,, Statistics, Purdue University
2013	MSC, Mathematical Statistics, Purdue University
2007	GRADUATE CERTIFICATE, Hopkins-Nanjing Center, Johns Hopkins SAIS
2000	BSC, Economics, University of Pittsburgh

Experience

2015-2016	Research Technician, Northeastern University
2010-2015	Research Assistant, Purdue University
2009-2010	Policy Coordinator, USITO
2007-2009	Managing Director, Jobdou Recruitment Services Ltd.
2005	Intern, Lenovo

Grants, honors & awards

2016	Montreal Causal Inference and Genetics Workshop Travel Award
2015	US HUPO Best Poster Award
2011	Boren Graduate Fellowship
2010	Purdue Doctoral Fellowship
2004	Boren Undergraduate Fellowship

Publications

Robert Ness, Karen Sachs, Olga Vitek. From correlation to causality: statistical approaches to learning regulatory relationships in large-scale biomolecular investigations", *Journal of Proteome Research* 15.3 (2016): 683-690.

Associations

American Statistical Association
International Society for Computational Biology