

Exploring Bayesian Networks: Replicating Causal Protein Signalling Study

GREGORY TOMY

Applied Mathematics, University of Colorado Boulder

ABSTRACT

This study replicates Sachs et al.'s (2005) "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data" using a Bayesian networks (BN). We explore the fundamental concepts of BNs, emphasizing their role in causal inference. We successfully reconstructed the 17 arcs reported in the original paper using the Tabu search algorithm with MBDe score. Five false positives were identified in our analysis, indicating potential overinclusiveness of the model. We discuss the implications of these findings and highlight the practical applications of BNs in unraveling causal relationships in complex biological systems. Additionally, we address the replication issues encountered during the project.

1. Introduction

Bayesian networks (BNs) have emerged as a powerful tool in probabilistic modeling, enabling us to unravel complex relationships and infer causal dependencies within a wide range of domains. These networks provide a formal framework for reasoning under uncertainty. Their applications span various fields, including biology, medicine, decision-making, risk analysis, and artificial intelligence, making them an invaluable tool for researchers and practitioners alike.

The present project aims to explore Bayesian networks by replicating the seminal work of Sachs et al. (2005) on the automated derivation of causal protein signalling networks. We aim to deepen our understanding of the underlying principles of BNs, showcase their applications, and illustrate potential implications of causal reasoning using BNs. We introduce BNs through a toy example, presenting an overview of their fundamental concepts and illustrating their utility. Additionally, we briefly discuss the concept of causal reasoning and its association with BNs. Following this groundwork, we provide a concise summary of the Sachs paper and then proceed to conduct the analysis. In subsequent sections, we present the results and showcase the process of inference within the BN framework. We discuss our findings and the knowledge gleaned, underscoring the practical implications. Finally, we address the replication issues encountered during the project.

2. Bayesian Networks: A Quick Tutorial

Bayesian Networks are probabilistic graphical models that offer a succinct and intuitive representation of the intricate relationships among a set of random variables. BNs

utilize a directed acyclic graph (DAG) structure, where nodes correspond to random variables, and edges represent the dependency relationship between variables. By leveraging Bayes' Theorem, BNs enable the specification of the conditional probability distribution for each node, given the states of its parent nodes. Consequently, these networks serve as a powerful tool for modeling complex and uncertain systems, allowing for predictions and informed decision-making based on available evidence.

a. Building blocks

The present section explains Bayesian networks through a toy example from Jensen and Nielsen (2007).

Suppose you find one morning that your car does not start. The starter is turning, but nothing happens. Your problem could be caused by a number of factors. You can hear the starter roll, so the battery must be charged. Thus, you conclude that the most likely causes are that the fuel was stolen overnight or that the spark plugs are dirty.

We build a Bayesian Network that depicts the causal relationships between events in Fig 1.

1) NODES

In BNs, nodes represent variables or events we are interested in modeling. Each node can take a finite set of mutually exclusive states. In the toy model, we have nodes and states *Fuel?* {yes, no}, *Clean Spark Plugs?* {yes, no}, *Fuel Meter Reading* {full, half, empty}, and *Start?* {yes, no}.

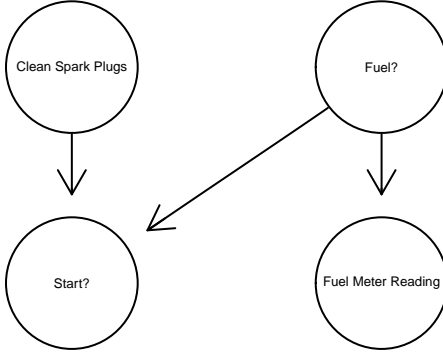


FIG. 1. Car Start Problem

TABLE 1. $P(FM|Fu)$: probability of fuel meter reading (FM) given fuel (Fu)

	Fu= yes	Fu= no
Full	0.39	0.001
Half	0.60	0.001
Empty	0.01	0.998

TABLE 2. $P(S|SP, Fu = Yes)$: probability of start (S) given if clean spark plug is dirty (SP) when there is fuel.

	SP = yes	SP = no
Yes	0.99	0.01
No	0.01	0.99

2) EDGES

Edges in a BN represent the dependencies between variables. An edge from *Fuel?* \rightarrow *Start* indicates that former has an effect on the later. Using the language of causality, we say that the state of *Fuel?* has a causal impact on the state of *Start*.

We call *Fuel?* the *parent* node of *Start* and *Start* the *child* node of *Fuel?*.

3) CONDITIONAL PROBABILITY TABLES

The strength of the relationship between nodes in a BN is represented by conditional probability tables (CPTs). CPTs describe the probability distribution of each node given the state or value of its parent nodes. Let $P(\text{Fuel?}) = (0.98, 0.02)$ and $P(\text{Clean Spark Plugs?}) = (0.96, 0.04)$. The remaining CPTs given in Table 1, 2, and 3.

b. Chain Rule for Bayesian Networks.

The chain rule for Bayesian Networks states that the joint probability of all the nodes in the network can be computed by multiplying the conditional probabilities of each node given its parent nodes. Although the joint probability distribution models the system completely, as the

TABLE 3. Car not starting.

	Fu	SP
Yes	0.71	0.42
No	0.29	0.58

number of variables increases, calculating the joint probability directly becomes computationally challenging. The chain rule allows us to break this calculation into a product of smaller, more manageable conditional probabilities. Furthermore, by multiplying the appropriate updated conditional probabilities, the chain rule allows for the propagation of fresh evidence into the network. In addition, the chain rule provides an efficient, modular representation of the network.

$$P(Fu, FM, SP, ST) = P(Fu)P(SP)P(FM|Fu)P(ST|Fu, SP)$$

c. Querying the Network

With all the essential components in place, we can now query the network. BNs are used for calculating new probabilities when presented with new information. In our case, we know that the car doesn't start. We say that the node *Start?* has been *instantiated* to $S = No$. When evidence is introduced to a node in a BN, the corresponding CPT for that node is updated to reflect the new information. These updates consider the marginal probabilities of the node given the evidence at hand. Furthermore, CPTs for other nodes directly influenced by the instantiated node may also need to be updated. To achieve this, the updated probabilities are propagated through the network, using the rules of probability theory.

Table 3 shows how the evidence $S = No$ has changed our perceptions of *Fu* and *SP*. Both the belief that we don't have enough gasoline and the belief that the spark plug is dirty have increased. This is consistent with intuition; we know there is a problem, but we lack sufficient evidence to pinpoint it. Continuing the story:

To find out the issue, you look at the fuel meter. It shows half full, so you decide to clean the spark plugs

In the language of BNs, we have instantiated another node, $FM = \text{"half"}$. Table 4 shows how the additional evidence, $FM = \text{"half"}$ has updated our beliefs regarding *Fu* and *SP*. After checking the fuel meter reading, we are sufficiently certain that it spark plugs are the issue. Note, however, that the BN allows for some probability that it is another issue entirely.

3. Causal Reasoning.

The ability to reason causally is a fundamental aspect of human cognition that we often taken for granted. Indeed,

TABLE 4. After checking fuel meter.

	Fu	SP
Yes	1	0.2
No	0	0.8

the above example may appear trivial. However, replicating this kind of reasoning in a computer requires a system to represent the problem and perform inference using these representations. This is a key strength of Bayesian Networks. The conclusion “*The fuel meter is at half, thus lack of fuel does not seem to be the reason for the car not starting so most likely the spark plugs are not clean*” is arrived at computationally using BNs.

Bayesian networks, in and of themselves, do not imply causal links. However, BNs can be viewed as a type of causal network that contains probabilistic information regarding variable interactions. The directed edges can be interpreted as causal relationships, and the conditional probabilities associated with each node can be interpreted as quantifying the causal relationship between parent nodes and the child nodes. BNs must conform to *d-separation* properties in order to accurately represent causality. When two nodes are *d-separated*, all pathways between them are blocked, and information entering one has no influence on the other.

a. Ladder of Causation

The ladder of causation is a framework proposed by Judea Pearl for understanding different levels of causation. It consists of three rungs, each representing a distinct perspective on causality. The first rung of the ladder, *association*, captures the correlations of variables without explicitly establishing causal relationships. While observing associations can be informative, they do not provide a clear understanding of cause and effect. The second rung, *intervention*, enables us to establish causal directionality. By manipulating one variable and observing its impact on another, we gain insights into the cause-effect relationship between them. This forms the basis for experimental studies and interventions, enabling researchers to actively test hypotheses and make causal claims. The highest rung on the ladder, *counterfactuals*, adds another layer of complexity to causal reasoning. Here, we engage in reasoning about what would have happened under alternative conditions. Counterfactual thinking enables us to assess the causal effect of an intervention without needing to actually do the intervention. By exploring counterfactual scenarios, we gain a deeper understanding of the causal mechanisms at play and can evaluate the impact of interventions or policy changes.(Pearl and Mackenzie 2018). Table 5 presents the ladder of causation with associated causal queries from the toy example.

TABLE 5. Ladder of Causation

Rung	CausalQuery
Counterfactual	How are fuel meter reading and fuel related?
Intervention	How would fuel change if I moved the fuel meter reading?
Association	What would have happened if the fuel level had been low, and the sparks plug clean?

4. Replication of Sachs et al.

Bayesian Networks provide a useful tool for the analysis of biological data due to their capacity to model complex relationships among variables, identify causal relationships, and make predictions. In biological research, BNs have been applied to a diverse range of problems, from elucidating genetic interactions to modeling cellular signaling pathways. The latter is explored in this paper. Signaling pathways are complex networks of proteins that transmit signals within cells, and their dysregulation can lead to many diseases, including cancer. BNs have been used to model these pathways and identify key proteins and interactions critical for the pathway’s function (Mukherjee and Speed 2008; Ciaccio et al. 2010). Signaling pathways data is ideal for BNs due to its high-dimensional and complex nature, making it difficult to interpret using traditional statistical approach. BNs can model nonlinear and dynamic relationships among variables and can also handle incomplete or noisy data, making them robust in the face of data quality issues that are commonly encountered in biological research. Furthermore, BNs allow for the integration of prior knowledge into the analysis, such as known interactions among proteins, which can improve accuracy and interpretability of the results.

An excellent illustration of utilizing protein-signaling data is presented in Sachs et al. (2005). BNs were used to automate derivation of causal influences in cellular signaling networks. The authors simultaneously measured multiple phosphorylated molecules in thousands of human single cells. The molecules were treated with generic and specific molecular interventions. The former was used to generate the network skeleton, whilst the latter was used to interpret causal influence relationships.

a. Summary of Sachs et al. (2005)

The analysis in Sachs et al. (2005) is summarized below.

1. The authors employed a single-cell approach to measure multiple protein states, which allowed for the collection of thousands of data points and eliminated population-averaging effects.

2. Interventional assays generated hundreds of data points per intervention, increasing the ability to infer causality.
3. Bayesian network inference was performed on the data sets, demonstrating the critical importance of interventions for effective inference, particularly in establishing directionality of connections.
4. The resulting BN model was compared to established signaling pathways in the literature, revealing the successful reverse-engineering and inference of a classical signaling network’s basic structure.

5. Data

Two data sets were used in the project. The first observation-only dataset was obtained from Sachs et al. (2005) supplementary materials. The observational data served as the basis for exploratory and preliminary analysis. There were 852 observations for the 11 phosphorylated molecules. The second data set was obtained through Scutari and Denis (2021)’s analysis. It contained 5400 data points which included the observation-only data in addition to stimulatory and inhibitory intervention data. A 12th column, *INT*, recorded the intervened molecules. The use of these different sources is explained in the *Issues with Replication* section.

a. Data exploration

Molecule concentrations were found to exhibit significant skewness (Fig 2), as well as nonlinear dependence relationships (Fig 3). Furthermore, all 11 molecules had outliers in the data.

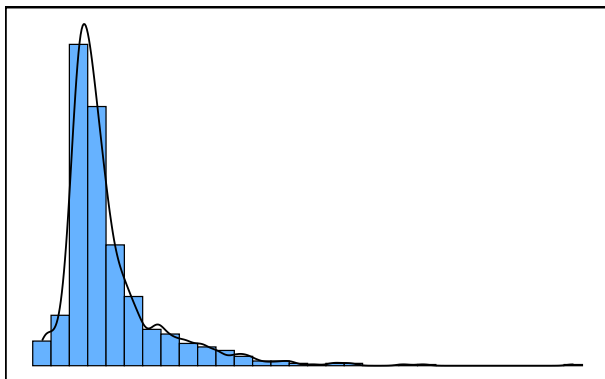


FIG. 2. PKA concentration showing a skewed distribution

b. Data preparation

As described by Friedman (2004), molecule concentrations are assumed to be normally distributed and Gaussian Bayesian Networks (GBN) are the standard for modeling signaling networks. However, data exploration revealed

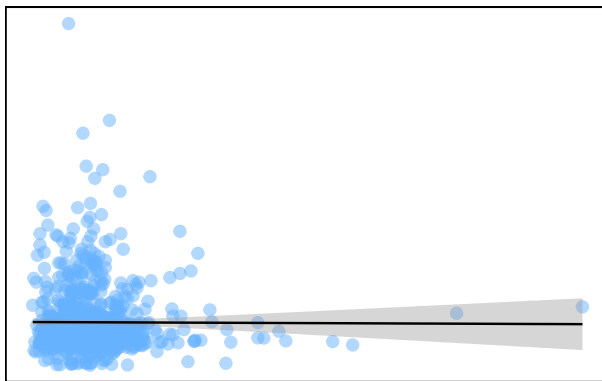


FIG. 3. Nonlinear dependence relationship between Raf and PK.

heavily skewed molecule concentration and nonlinear dependence relationships. To address this issue, Sachs et al. (2005) discretized their data. Discretization of continuous variables can be beneficial in reducing the impact of skewness and modeling nonlinear relationships (Koller and Friedman 2009).

The data was discretized into three concentrations (*low*, *medium*, *high*) using Hartemink’s information-preserving algorithm (Sachs et al. 2006). The Hartemink algorithm can handle nonlinear relationships and heavy-tailed distributions by utilizing a non-uniform discretization scheme that assigns more categories to the tails of the distribution. This allows for better resolution in these regions and effectively captures the relevant features of the data (Hartemink 2001; Scutari and Denis 2021).

The second complete data set had already been pre-processed and discretized, and no changes were made.

6. Structural Learning

Learning the structure of a BN from data involves identifying the conditional dependence relationships among the variables, which are achieved through two main methods: constraint-based and score-based methods. Constraint-based methods establish a set of conditional independence statements that hold true for the given data, and then construct a network with d-separation attributes that correspond to these properties. On the other hand, score-based approaches generate a set of candidate Bayesian networks, assign each candidate a score, and select the candidate with the highest score (Jensen and Nielsen 2007). In general, score-based methods are better suited for discrete Bayesian networks due to the sparsity of the data and the large number of possible network structures (Friedman et al. 1997). Following Sachs et al. (2006) thesis paper, score-based learning with the Bayesian Dirichlet equivalent (Bde) score was utilized in this project.

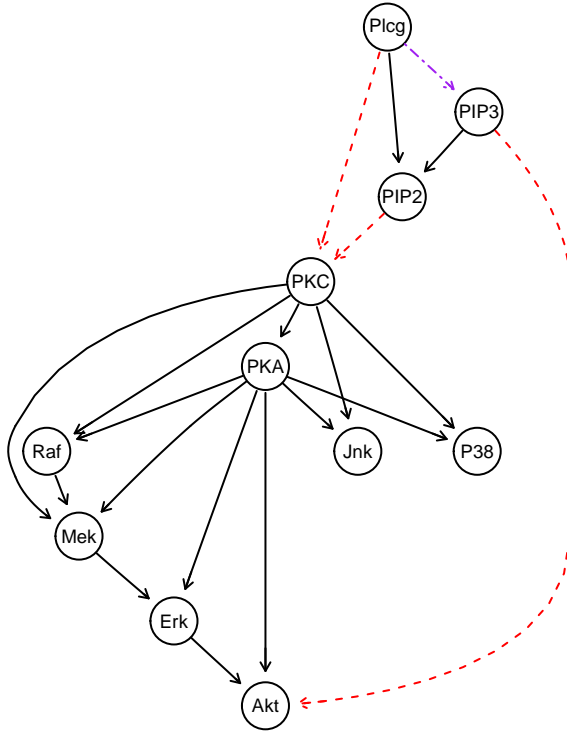


FIG. 4. Target BN from Sachs et al. (2005). The red dotted arcs denote the arcs not accounted for by Sachs' model. The purple dotted arc denotes a reversed arc.

a. Hill Climbing Algorithm (HC)

HC is a popular score-based algorithm and was the primary algorithm used for structural learning in the project. The algorithm starts with an empty network and iteratively explores different network structures by adding or removing arcs while evaluating their corresponding scores with the aim to maximize the chosen scoring function (Koller and Friedman 2009).

b. Tabu Search Algorithm (Tabu)

Tabu is an extension of the HC algorithm. Tabu search uses a memory-based strategy to intensify the search in promising regions of the search space, while also diversifying the search by exploring different areas of the search space (Russell 2010).

7. Analysis

a. Observational data

Observational data was first modeled utilizing the hill climbing algorithm. A bootstrap of 1000 candidate networks was used to construct an average network (M1). The observational data was also modeled using random

sampling with the Uniform Random Acyclic Digraphs algorithm (Melançon and Philippe 2004). 1000 random networks were generated from a uniform distribution over the set of all possible graphs. Each of these were then used as a starting point for hill climbing algorithms that were used to construct an average network (M2). For both models, a significance threshold of 0.9 was applied to include only the strongest arcs in the network.

b. Full data

The interventional data was used by Sachs et al. (2005) to specify causal relationships in the network. We first modeled the data by considering *INT* as its own variable and forced all the nodes in the network to be dependent on it (M3). Next we modeled the data by blocking any arcs from the network to *INT*. This had the advantage of letting the algorithm decide which arcs connecting *INT* to the other nodes should be included in the network (M4). Both models were implemented with the hill-climbing algorithm.

The final model (M5) used the random sampling method in M2 but with two changes. Following the analysis done by Scutari and Denis (2021), we incorporated the impact of the interventions into the score components associated with each node. To achieve this, the *Modified Bayesian Dirichlet equivalent (MBDe)* score was utilized. When an intervention is applied to a node, its value is fixed, and is no longer determined by the values of its parents in the network. The MBDe score adjusts for this by ignoring the effects of the parents on intervened nodes (Cooper and Yoo 2013). Furthermore, *Tabu search* algorithm was implemented to avoid getting stuck in a local optima.

8. Results

M1 results were promising. As Fig 5 shows, the BN modeled the major protein interaction clusters. However, there were still 7 arcs missing. Crucially, the BN could not establish directionality for any of the arcs. M2 returned the exact same BN as the bootstrap network. Nevertheless, the results also highlighted the limited extent of information that could be extracted from the observational data set alone.

M3 and M4 modeled more of the structural characteristics of the target network. The relationships *PKA-Erk-Akt* were modeled correctly in the M3 (Fig 6), but not in M4 (Fig 7). *P38-PKC-Jnk* and *PIP3-PIP2-Plcg* clusters were learned correctly in both BNs. Although M3 captured the causal direction *Raf* \rightarrow *Mek* correctly, both BNs missed the *Raf-Mek-Erk-Akt* cascade.

M5 successfully identified all the arcs from the target network. The three arcs missed by Sachs et al. (2005) were also missed by M5. There were no false negatives, however, the model did return 5 false positives. Further

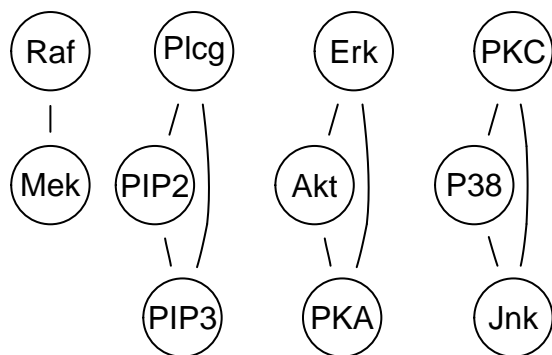


FIG. 5. M1 averaged network from bootstrap sample

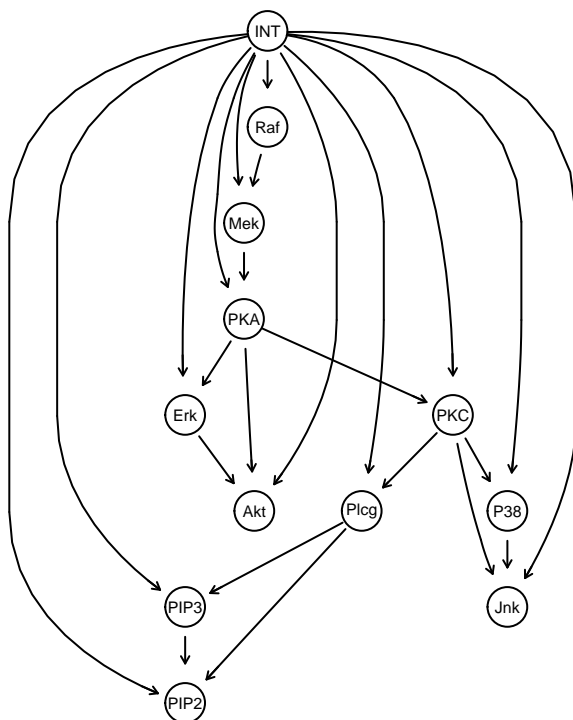


FIG. 6. M3 network learned using forced arcs.

increase in the significance resulted in the breakdown of the network structure. This result highlighted the superior performance of the simulated annealing algorithm used by Sachs et al. (2006).

9. Inference

Bayesian networks can be queried to make probabilistic predictions about the network variables. The model should correctly infer causal influences for molecules that were not directly intervened with. Sachs et al. (2005) investigate two such queries in their study. The model presented the arc $Erk \rightarrow Akt$ despite Erk not being directly perturbed. As a result, changing Erk should have an effect

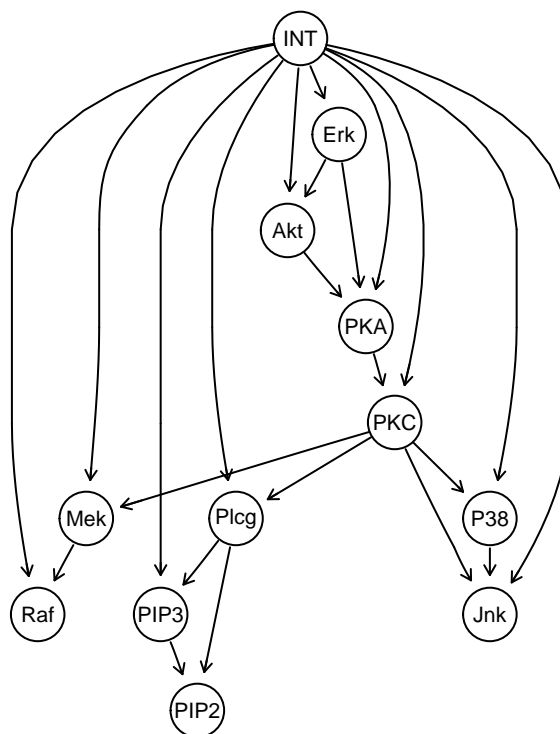


FIG. 7. M4 network learned using blocked arcs

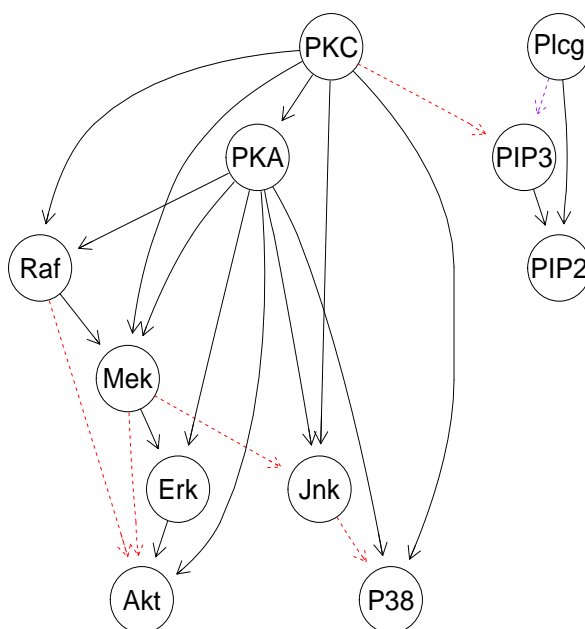


FIG. 8. M5 network. Red dashed arcs denote the false positives. The purple dotted arc denotes a reversed arc.

on *Akt*. Similarly, despite being correlated, the model did not present an arc from *Erk* to *PKA*. Perturbing the former should have no effect on the latter. Sachs et al. (2005)

verified these predictions experimentally. In this project, the model was queried by instantiating *Erk* to *high*. As expected, instantiating *Erk* did effect the probability distribution of *Akt* (Fig 9), but did not effect *PKA* (Fig 10).

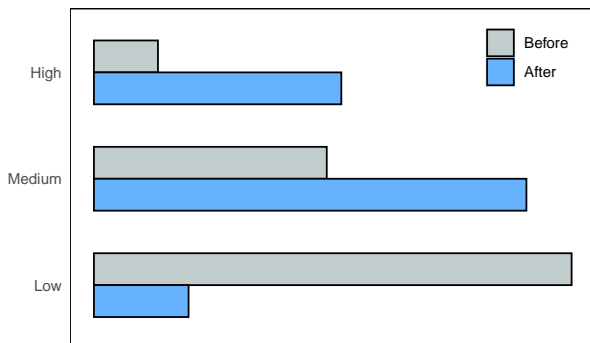


FIG. 9. Probability distribution of Akt before and after instantiating Erk

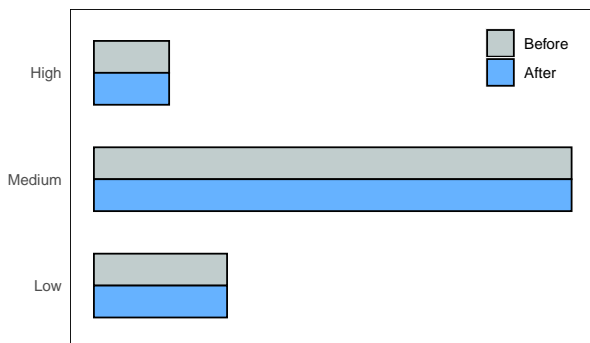


FIG. 10. Probability distribution of PKA before and after instantiating Erk

10. Discussion

In this project, we successfully replicated the Bayesian network proposed by Sachs et al. (2005). The analysis showed the effectiveness of BNs in modeling complex causal relationships among proteins involved in cellular processes. The BN was constructed without any prior knowledge of pathway connectivity, highlighting the usefulness of this approach in automatically learning novel causal relationships. One fundamental aspect of BNs is the probabilistic nature of the analysis. Effectively inferring relationships and making accurate predictions requires a sufficient amount of data. Therefore, it is crucial to ensure an adequate sample size and multiple observations to obtain reliable results. Observational data, however, only takes us to the first rung of the causal ladder. As seen in this project, experimental interventional data was crucial to get to the second rung and determine

causal directions. The BN also detected causal relations in molecules that were not directly intervened, such as the well-known signal pathway from *Raf* to *Mek* (Sachs et al. 2005). A major strength displayed by the BN was its ability to detect both direct and indirect causal relationships. The *PKC* \rightarrow *Jnk* pathway modeled involves two intermediate proteins that were not measured, yet the network correctly detected the causal direction (Sachs et al. 2005). This is particularly useful when the data is not exhaustive. Furthermore, the BN dismissed connections that were already explained by existing arcs. Consider again the pathway *Raf* \rightarrow *Mek* \rightarrow *Erk*. Although *Erk* is dependent on *Raf*, there was no direct arc present between them since the pathway through *Mek* explained the dependence. In contrast, the BN modeled both a serial pathway *PKC* \rightarrow *Raf* \rightarrow *Mek* and a direct pathway *PKC* \rightarrow *Mek*. As Sachs et al. (2005) explain, *PKC* is known to affect *Mek* through two pathways. This demonstrated the ability of the BN to model complex casual relationships by detecting multiple causal paths. The BN also detected lesser-known arcs that were experimentally verified by Sachs et al. (2005). This ability may prove invaluable to researchers by providing important clues on what areas to investigate further. We showed how counterfactuals can be framed as queries to the network, enabling us to ascend to the top rung of the causal ladder. By formulating these counterfactual queries and propagating them through the network, we can obtain probabilistic estimates of outcomes of interest. This approach provides a powerful means to explore the causal structure of the system and understand the potential consequences of specific interventions or changes in variables.

One drawback of BNs is their inability to model cyclic arcs. The three cyclic arcs missed by Sachs et al. (2005) were also missed by our model. Several models have been proposed to address this issue, for example *Dynamic Bayesian Networks* by Murphy (2002), which allow for cyclic dependencies and temporal data. It is also important to remember that whilst BNs excel at representing associations and correlations they have limitations when it comes to inferring causality. While causal relationships can be inferred under certain conditions, like in the case of *PKA* \rightarrow *Erk* \rightarrow *Akt* pathway, BNs themselves do not provide definite evidence of causality. Nonetheless, coupled with interventional experimental design, they are a valuable tool for casual inference.

11. Issues with Replication

Replication is a cornerstone of scientific research, providing essential means to validate and build on existing findings. In this section, we discuss the replication issues encountered while attempting to replicate the study

by Sachs et al. (2005). There were significant discrepancies between data sets provided in the supplementary materials of the paper and the data sets mentioned. The full data set with 5400 data points cited in the paper appears nowhere in the data sets provided. To overcome this hurdle, we sought out an alternative source and obtained the complete data set from the analysis by Scutari and Denis (2021). Scutari conducted a similar analysis replicating the Sachs study and relied on Sachs' data sets provided with the original thesis paper. We were unable to find these data sets. Consequently, the project efforts heavily relied on Prof. Scutari's work and the data sourced from his analysis. Taking the cue from Scutari, we also used Sachs' thesis paper to inform our analysis. There was also a lack of comprehensive documentation and code, both in the Sachs et al. (2005) paper and Sachs et al. (2006) thesis paper. Here too, we relied on the analysis by Scutari and Denis (2021).

The replication challenges encountered in this project highlight a significant facet of the broader replication crisis in science. The inadequacy of documentation, presence of incomplete or inconsistent data sets, and restricted access to original sources collectively impede the replication of studies, thereby jeopardizing the accuracy and reproducibility of findings. One possible approach for mitigating these challenges involves the adoption of scripting languages that enable researchers to document their analysis procedures, algorithms, and code. Implementation of transparent research practices, such as the sharing of comprehensive data sets, well-documented methodologies, and utilization of open source practices, can further serve to alleviate these issues.

APPENDIX

a. Software and programming language

The data analysis for this project was conducted using the R programming language (v.4.2). The `bnlearn` package (v.4.8) was used for learning the graphical structure of the BN (Scutari 2009). Exact inference was implemented using the `gRain` (v.1.3) package (Højsgaard 2012). The `tidyverse` package (v.2.0) was utilized for data processing and visualization (Wickham et al. 2019).

b. Code

The code for this project and study along with datasets are available at <https://github.com/GregoryTomy/STAT5630-CompBayes.git>

c. Figures and charts

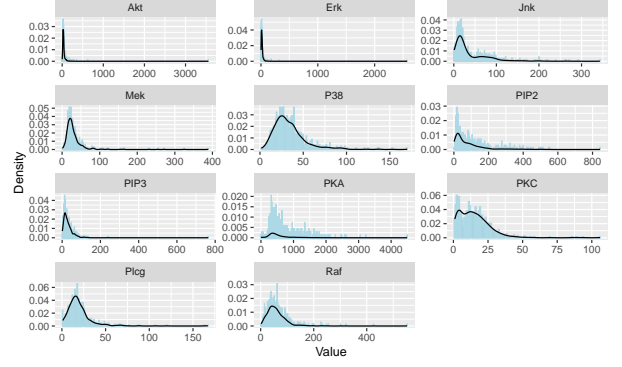


FIG. A1. Concentration histograms and density plots for the 11 phosphorylated molecules.

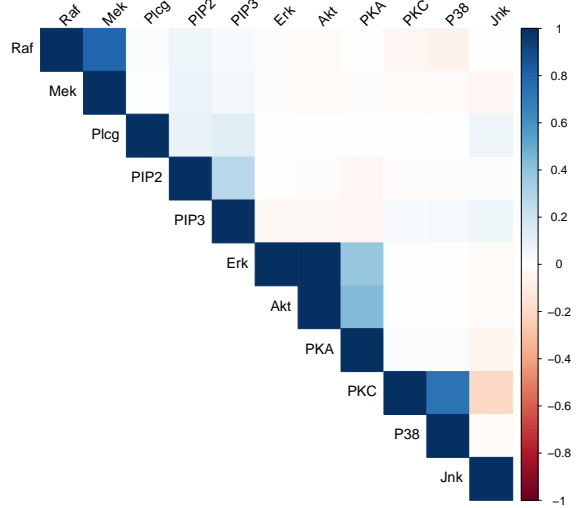


FIG. A2. Correlations matrix for the 11 phosphorylated molecules.

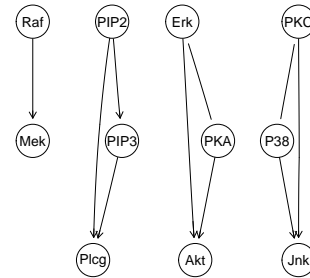


FIG. A3. M2 model network using random sampling

References

- Ciaccio, M. F., J. P. Wagner, C.-P. Chuu, D. A. Lauffenburger, and R. B. Jones, 2010: Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nature methods*, **7**, 148–155.
- Cooper, G. F., and C. Yoo, 2013: Causal discovery from a mixture of experimental and observational data. *arXiv preprint arXiv:1301.6686*,.
- Friedman, N., 2004: Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
- , D. Geiger, and M. Goldszmidt, 1997: Bayesian network classifiers. *Machine learning*, **29**, 131–163.
- Hartemink, A. J., 2001: Principled computational methods for the validation discovery of genetic regulatory networks. Massachusetts Institute of Technology,.
- Højsgaard, S., 2012: Bayesian networks in r with the gRain package. *Journal of Statistical Software*, **46**, 1–26.
- Jensen, F. V., and T. D. Nielsen, 2007: *Bayesian networks and decision graphs*. Springer,.
- Koller, D., and N. Friedman, 2009: *Probabilistic graphical models: Principles and techniques*. MIT press,.
- Melançon, G., and F. Philippe, 2004: Generating connected acyclic digraphs uniformly at random. *Information Processing Letters*, **90**, 209–213.
- Mukherjee, S., and T. P. Speed, 2008: Network inference using informative priors. *Proceedings of the National Academy of Sciences*, **105**, 14313–14318.
- Murphy, K. P., 2002: *Dynamic bayesian networks: Representation, inference and learning*. University of California, Berkeley,.
- Pearl, J., and D. Mackenzie, 2018: *The book of why: The new science of cause and effect*. Basic books,.
- Russell, S. J., 2010: *Artificial intelligence a modern approach*. Pearson Education, Inc.,.
- Sachs, K., and Coauthors, 2006: Bayesian network models of biological signaling pathways. Massachusetts Institute of Technology,.
- , O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan, 2005: Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.
- Scutari, M., 2009: Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*,.
- , and J.-B. Denis, 2021: *Bayesian networks: With examples in r*. CRC press,.
- Wickham, H., and Coauthors, 2019: Welcome to the tidyverse. *Journal of open source software*, **4**, 1686.