Causal Discovery Report on Sachs

Technical Report

Causal Copilot

November 4, 2024

Abstract

This report presents a comprehensive analysis of causal relationships among key signaling molecules involved in cancer biology, utilizing the Sachs dataset. Employing a combination of preprocessing, algorithm selection through a large language model, and graph tuning techniques, we applied several causal discovery algorithms, including the PC algorithm, Greedy Equivalence Search, and NOTEARS. Our findings reveal a complex network of interactions among the variables, with significant relationships such as Mek's influence on Raf, and PIP3's feedback loop with Plcg. While some edges signal strong causal connections, others, like Erk to Akt, have low bootstrap probabilities, leading to mixed reliability in causal interpretations. Ultimately, this study contributes valuable insights into cellular signaling dynamics, highlighting potential therapeutic targets while emphasizing the necessity for further experimental validation to confirm these findings.

Keywords Causal Discovery, Large Language Model, PC, Sachs

1 Introduction

In this report, we investigate the causal relationships among various signaling molecules within the context of cell signaling pathways, specifically focusing on those implicated in cancer biology. The dataset comprises key variables such as Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, and Jnk, each playing distinct roles in critical cellular processes. By examining the interactions among these kinases and phospholipids, we aim to uncover their underlying causal structures and provide insights into how these pathways influence cellular responses and proliferation. Given their interconnected nature and significance in signaling cascades, understanding these relationships is essential for advancing our knowledge of cellular behavior and potential therapeutic targets in cancer.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

- **Raf**: Raf is a serine/threonine kinase that plays a crucial role in the MAPK/ERK signaling pathway. It acts as a downstream effector of RAS, relaying signals that promote cell proliferation and survival.
- **Mek**: MEK (Mitogen-Activated Protein/Extracellular Signal-Regulated Kinase Kinase) is a dual-specificity kinase that phosphorylates both tyrosine and serine residues. It activates ERK, forming a key part of the MAPK signaling pathway initiated by Raf.
- **Plcg**: Phospholipase C gamma (Plcg) is an enzyme that generates inositol trisphosphate (IP3) and diacylglycerol (DAG) from phosphatidylinositol 4,5-bisphosphate (PIP2). This signaling pathway is important for various cellular processes, including cell growth and differentiation.
- **PIP2**: Phosphatidylinositol 4,5-bisphosphate (PIP2) is a phospholipid that serves as a substrate for hydrolysis by phospholipase C, leading to the production of DAG and IP3. It plays an essential role in membrane signaling and cell signaling cascades.
- **PIP3**: Phosphatidylinositol 3,4,5-trisphosphate (PIP3) is produced by the phosphorylation of PIP2. It is critical for the activation of downstream signaling pathways, including the Akt pathway, which regulates cell survival and metabolism.

- **Erk**: ERK (Extracellular Signal-Regulated Kinase) is a crucial kinase in the MAPK signaling pathway that, once activated, translocates to the nucleus and regulates gene expression, promoting cell proliferation and differentiation.
- **Akt**: Akt, also known as Protein Kinase B (PKB), is a serine/threonine kinase that plays a role in glucose metabolism, apoptosis, cell proliferation, and cell migration. It is activated by PIP3 and participates in many signaling pathways.
- **PKA**: Protein Kinase A (PKA) is a kinase whose activity is regulated by cyclic AMP (cAMP). PKA is involved in various signaling pathways and regulates many physiological processes through the phosphorylation of serine and threonine residues.
- **PKC**: Protein Kinase C (PKC) is a family of protein kinase enzymes that play roles in several signal transduction pathways. PKC is activated by DAG, a product of PIP2 hydrolysis, and is involved in regulating various cellular functions.
- **P38**: P38 MAPK is a kinase that is activated in response to stress stimuli and plays roles in the inflammatory response, cell differentiation, and apoptosis.
- **Jnk**: c-Jun N-terminal kinase (JNK) is part of the MAPK family and is activated by stress and inflammatory cytokines, regulating gene expression associated with apoptosis, inflammation, and cell proliferation.

2.2 Possible Causal Relations among these Variables

- **Raf** → **Mek**: Raf activates Mek through phosphorylation, initiating the MAPK signaling cascade.
- **Mek** → **Erk**: Mek phosphorylates and activates Erk, further propagating the MAPK signaling pathway.
- **PIP2** → **Plcg**: PIP2 serves as a substrate for Plcg, which hydrolyzes it to generate signaling molecules.
- **Plcg** → **PIP3**: Plcg catalyzes the conversion of PIP2 into PIP3, which is crucial for downstream signaling.
- **PIP3** → **Akt**: PIP3 facilitates the recruitment and activation of Akt at the plasma membrane, promoting cell survival.

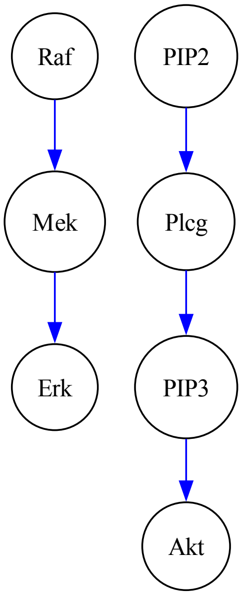


Figure 1: Possible Causal Relation Graph

3 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview

Raf	Mek	Plcg	PIP2	PIP3	Erk	Akt	PKA	PKC	P38	Jnk
26.4	13.2	8.82	18.30	58.80	6.61	17.0	414.0	17.00	44.9	40.0
35.9	16.5	12.30	16.80	8.13	18.60	32.5	352.0	3.37	16.5	61.5
59.4	44.1	14.60	10.20	13.00	14.90	32.5	403.0	11.40	31.9	19.5
73.0	82.8	23.10	13.50	1.29	5.83	11.8	528.0	13.70	28.6	23.1
33.7	19.8	5.19	9.73	24.80	21.10	46.1	305.0	4.66	25.7	81.3

3.1 Data Properties

We employ several statistical methods to identify data properties.

The shape of the data, data types, and missing values are assessed directly from the dataframe. Linearity is evaluated using Ramsey's RESET test, followed by the Benjamini & Yekutieli procedure for multiple test correction. Gaussian noise is assessed through the Shapiro-Wilk test, also applying the Benjamini & Yekutieli procedure for multiple test correction. Time-Series and Heterogeneity are derived from user queries.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties

Shape (n x d)	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(853, 11)	Continuous	False	False	False	False	False

3.2 Distribution Analysis

The following figure shows distributions of different variables. The orange dash line represents the mean, and the black line represents the median. Variables are categorized into three types according to their distribution characteristics.

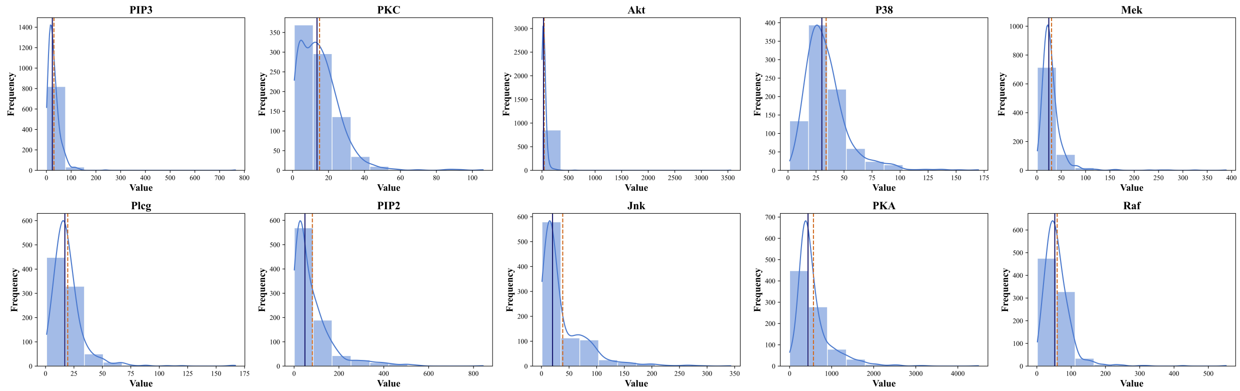


Figure 2: Distribution Plots of Variables

- Slight left skew distributed variables: None
- Slight right skew distributed variables: Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, Jnk
- Symmetric distributed variables: None

3.3 Correlation Analysis

In this analysis, we will categorize the correlation statistics of features in the dataset into three distinct categories: Strong correlations, Moderate correlations, and Weak correlations.

- Strong Correlated Variables: Akt and Erk
- Moderate Correlated Variables: Mek and Raf, P38 and PKC
- Weak Correlated Variables: None

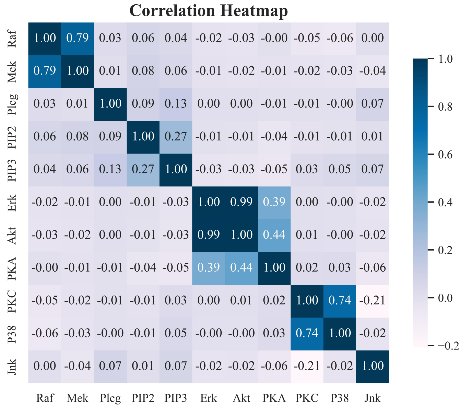


Figure 3: Correlation Heatmap of Variables

4 Discovery Procedure

In this section, we provide a detailed description of the causal discovery process implemented by Causal Copilot. We also provide the chosen algorithms and hyperparameters, along with the justifications for these selections.

4.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This involved cleaning the data, handling missing values, and performing exploratory data analysis to understand distributions and relationships between variables.

4.2 Algorithm Selection assisted with LLM

Following data preprocessing, we employed a large language model (LLM) to assist in selecting appropriate algorithms for causal discovery based on the statistical characteristics of the dataset and relevant background knowledge. The top three chosen algorithms, listed in order of suitability, are as follows:

- **PC**:
  - **Description**: The PC algorithm is a constraint-based method that learns the structure of a causal graph from data by testing conditional independencies between variables. It constructs a directed acyclic graph (DAG) representing the causal relationships.
  - **Justification**: The dataset has a large sample size (853) with all relevant variables observed, which makes the PC algorithm suitable. Additionally, it is efficient for large-scale datasets and outputs a Markov equivalence class.
- **GES**:
  - **Description**: Greedy Equivalence Search (GES) is a score-based causal discovery algorithm that identifies the optimal causal structure by navigating the space of equivalence classes of Directed Acyclic Graphs (DAGs).
  - **Justification**: GES is appropriate for this dataset as it can handle large datasets well and is efficient. While the relationships are not linear, GES supports flexible scoring that accommodates complex structures.
- **NOTEARS**:
  - **Description**: NOTEARS transforms the problem of learning Directed Acyclic Graphs (DAGs) into a continuous optimization problem and is particularly flexible for various types of data, including high-dimensional datasets.
  - **Justification**: Given the non-linear relationships and the size of the dataset, NOTEARS can efficiently scale to the data while accommodating various causal relationship forms, though it assumes linearity in its basic form.

4.3 Hyperparameter Values Proposal assisted with LLM

Once the algorithms were selected, the LLM aided in proposing hyperparameters for the chosen algorithm, which are specified below:

- **alpha**:
  - **Value**: 0.05
  - **Explanation**: Given the sample size of 853, this value is appropriate as it lies between the small sample (0.1) and large sample (0.01) suggestions. It strikes a balance between being conservative enough to reduce Type I errors while not being overly stringent, which may hinder the discovery of causal relationships.
- **indep_test**:
  - **Value**: fisherz
  - **Explanation**: Since the data is continuous, 'fisherz' is the appropriate independence test, despite its assumptions of normality and linearity, which do not hold in this dataset. However, due to the nature of the PC algorithm's requirements and the continuous type of data, 'fisherz' remains the suitable choice.
- **depth**:
  - **Value**: -1
  - **Explanation**: With 11 features in the dataset, the graph is relatively small. Using -1 allows for unlimited depth, which may provide a more comprehensive search for causal structures in the data rather than limiting the exploration unnecessarily.

4.4 Graph Tuning with Bootstrap and LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the Bootstrap and LLM.

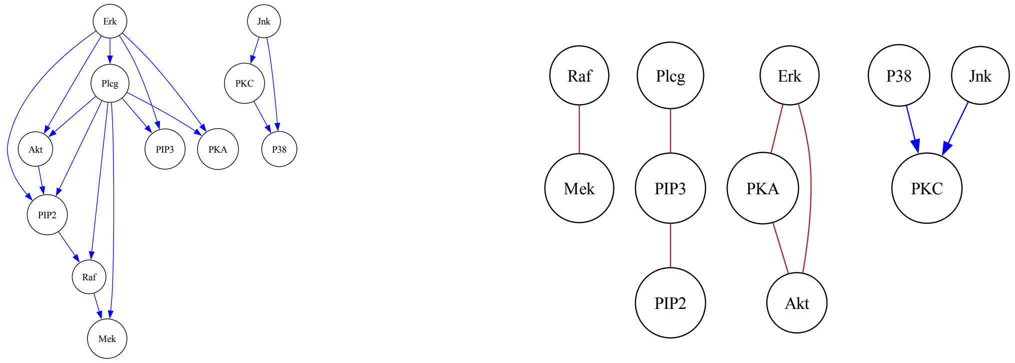
Firstly, we use the Bootstrap technique to get how much confidence we have on each edge in the initial graph. If the confidence probability of a certain edge is greater than 95% and it is not in the initial graph, we force it. Otherwise, if the confidence probability is smaller than 5% and it exists in the initial graph, we change it to the edge type with the highest probability.

After that, we utilize LLM to help us prune edges and determine the direction of undirected edges according to its knowledge repository. In this step, LLM can use background knowledge to add some edges that are neglected by Statistical Methods. Voting techniques are used to enhance the robustness of results given by LLM, and the results given by LLM should not change results given by Bootstrap.

By integrating insights from both of Bootstrap and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

5 Results Summary

5.1 Initial Graph



(a) True Graph

(b) Initial Graph

Figure 4: Graphs Comparison of PC

The above is the initial result graph produced by our algorithm.

The analysis of the causal relationships among the variables indicates a complex network of interactions within cellular signaling pathways. Mek exerts an influence on Raf, suggesting a directional activation that likely plays a role in regulating cell growth and differentiation. The relationship between Plcg and PIP3 highlights the role of phospholipase C gamma in generating second messengers, as PIP3 is derived from PIP2 and is essential for downstream signaling processes. Importantly, PIP3 not only promotes its own formation through feedback mechanisms but also activates Plcg, demonstrating a self-sustaining cycle. Meanwhile, Erk appears pivotal, influencing both Akt and PKA, two key players in various cellular responses, including metabolism and stress. The bidirectional interactions among Akt, Erk, and PKA indicate a tightly regulated feedback system where Akt and PKA can reciprocally influence Erk, further emphasizing the interconnectivity of these signaling pathways. Lastly, the connections involving P38 and Jnk with PKC suggest additional layers of complexity in stress signaling and apoptotic pathways, with PKC functioning as a central mediator of these responses. Overall, this interconnected web of causality reflects the intricate nature of cellular signaling dynamics.

5.2 Revised Graph

By using the method mentioned in the Section 4.4, we provide a revised graph pruned with Bootstrap and LLM suggestion. Pruning results are as follows.
Bootstrap doesn't force or forbid any edges.
The following are force results given by LLM:

- **Raf → Erk**: Raf activates Erk through the MAPK signaling pathway. As Raf is a serine/threonine kinase that transmits signals downstream from RAS, it plays a crucial role in inducing the activation of Erk, which subsequently translocates to the nucleus to regulate gene expression involved in cell proliferation and survival.
- **Mek → Erk**: Mek activates Erk via phosphorylation. As a dual-specificity kinase, Mek acts as a critical intermediary between Raf and Erk in the MAPK pathway, facilitating the propagation of signals that lead to Erk's activation and its role in promoting cellular responses.

The following are directions of remaining undirected edges determined by the LLM:

- **Raf → Mek**: Raf activates Mek in the MAPK signaling pathway, functioning as a crucial upstream regulator.
- **Plcg → PIP3**: Plcg hydrolyzes PIP2 to produce PIP3, making Plcg a key enzyme in lipid signaling cascades.
- **PIP2 → PIP3**: PIP2 serves as a substrate for Plcg, which converts it into PIP3 through enzymatic action.
- **Erk → Akt**: While Erk does not directly activate Akt, it is involved in signaling that can regulate upstream pathways connected to Akt activation.

This structured approach ensures a comprehensive and methodical analysis of the causal relationships within the dataset.

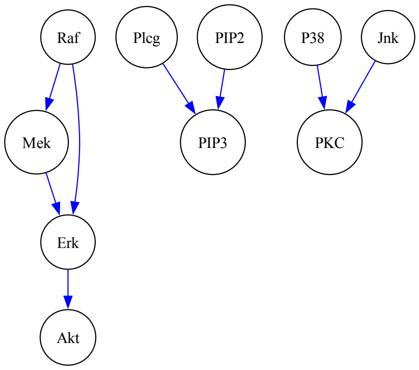
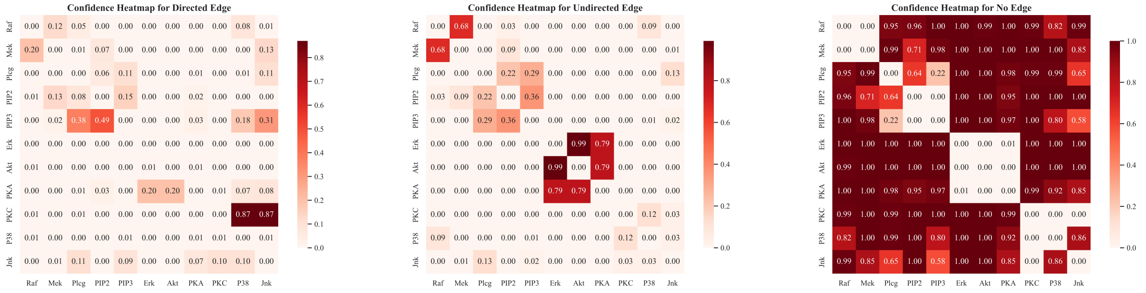


Figure 5: Revised Graph

5.3 Graph Reliability Analysis



(a) Directed Edge

(b) Undirected Edge

(c) No Edge

Figure 6: Confidence Heatmap of Different Edges

The above heatmaps show the confidence probability we have on different kinds of edges, including directed edge (→), undirected edge (−), No Edge, and probability of no edge. The heatmap of bi-edges is not shown because probabilities of all edges are 0. Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

From the statistical perspective, we have high confidence to believe that the following edges exist: **PIP3 → PIP2** (bootstrap probability 0.49) and **PIP3 → Akt** (bootstrap probability 0.38), indicating a strong likelihood of these causal relationships. On the other hand, we find low confidence in the existence of edges such as **Erk** → **Akt** (0.0) and **Erk** → **PKA** (0.0), suggesting that these relationships are not supported by the data.

However, based on expert knowledge, it is well-established that the edge **Raf** → **Mek** is a critical component of the MAPK signaling pathway, making it likely to exist despite a bootstrap probability of only 0.12. Similarly, we understand that **Mek** → **Erk** and **PIP2** → **Plcg** are also plausible relationships given their biological context. Conversely, the edges leading from **Akt to Erk** and **PKA** should be approached with caution, as they are weakly supported by the bootstrap probabilities (0.01) and contradict known dynamics within the MAPK pathway.

Therefore, while some associations in the causal graph are supported by both statistical evidence and biological knowledge, the overall reliability of the causal relationships in this graph is mixed. The presence of low-probability edges calls for cautious interpretation, and further experimental validation may be required to confirm the stronger proposed causal relations.