CAUSAL DISCOVERY REPORT ON SACHS

TECHNICAL REPORT

Causal Copilot

October 28, 2024

ABSTRACT

This report presents a causal discovery analysis of cellular signaling pathways using a comprehensive dataset of proteins and lipid molecules, specifically focusing on Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, and Jnk. We employed a mixed methodology that included systematic data preprocessing, algorithm selection, and hyperparameter tuning, facilitated by a large language model (LLM). The algorithms chosenPC, GES, and FCIwere justified based on the dataset's characteristics and underlying assumptions of causal relationships. Our results revealed intricate interdependencies within the signaling pathways, highlighting feedback loops and regulatory influences among key players, with varying confidence levels in the causal edges identified. Notably, while robust relationships were documented, discrepancies with prior biological knowledge suggest a need for reevaluation and experimental validation of certain causal assumptions. This analysis contributes to a deeper understanding of cell signaling dynamics and underscores the necessity of integrating experimental insights with computational methodologies in causal inference.

Keywords Causal Discovery, Large Language Model, PC, Sachs

1 Introduction

In this report, we will explore the intricate relationships among a set of variables that are integral to cellular signaling pathways, specifically focusing on proteins and lipid molecules involved in signal transduction. The dataset comprises key components such as Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, P38, and Jnk, which play pivotal roles in regulating various cellular processes, including cell division, differentiation, survival, and responses to stress. Understanding the potential causal relationships among these elements is crucial, as it allows us to elucidate the complex cascading signaling networks that govern cellular behavior. Additionally, the insights gained from this analysis will contribute to our knowledge of how these pathways interact under various biological conditions, highlighting the importance of both experimental data and temporal dynamics in establishing causality. As we delve into this causal discovery, we aim to uncover patterns and mechanisms that may further our understanding of signaling cascades in cellular biology.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

- **Raf**: A family of proteins that act as serine/threonine kinases. Raf is involved in the MAPK/ERK signaling pathway, which regulates cell division, differentiation, and survival. It functions as a crucial initiator in this pathway, responding to various growth factors and promoting downstream signaling.
- **Mek (MEK)**: Also known as MAPK/ERK kinase, MEK is a dual-specificity kinase that specifically activates ERK by phosphorylation. It serves as a key intermediary in the MAPK signaling cascade, linking upstream signaling molecules to the activation of ERK and subsequently influencing cellular responses.
- **Plcg (Phospholipase C gamma)**: An enzyme that generates inositol trisphosphate (IP3) and diacylglycerol (DAG) in response to receptor activation. Plcg plays a pivotal role in facilitating calcium signaling and various cellular responses like proliferation, differentiation, and apoptosis through its products.

- **PIP2 (Phosphatidylinositol 4,5-bisphosphate)**: A phospholipid present in the inner membrane of cells that acts as a substrate for phospholipase C, contributing to signal transduction pathways. PIP2 is crucial for the generation of secondary messengers that mediate various downstream signaling events.
- **PIP3 (Phosphatidylinositol 3,4,5-trisphosphate)**: A product of the phosphorylation of PIP2 by phosphoinositide 3-kinases (PI3K). PIP3 functions as a key secondary messenger in multiple signaling pathways, particularly in the activation of the serine/threonine kinase AKT, thereby regulating various cellular processes, including metabolism and cell survival.
- **Erk (Extracellular signal-regulated kinase)**: A central protein in the MAPK signaling pathway that, when activated, translocates to the nucleus to influence gene expression. Erk is critical in mediating signals from growth factors and plays a significant role in cell proliferation, differentiation, and survival.
- **Akt**: Also known as Protein Kinase B (PKB), Akt is a serine/threonine kinase essential for promoting cell survival and growth in response to growth factor signaling, particularly through the PIP3 pathway. It is involved in various processes, including glucose metabolism and cell cycle progression.
- **PKA (Protein Kinase A)**: A serine/threonine kinase regulated by cyclic AMP (cAMP). PKA phosphorylates a wide range of target proteins, thereby participating in numerous signaling pathways, including those involved in metabolism and gene expression.
- **PKC (Protein Kinase C)**: A family of serine/threonine kinases activated by diacylglycerol (DAG) and calcium ions. PKC plays diverse roles in cellular processes, including growth regulation, differentiation, and the response to extracellular signals.
- **P38**: Part of the MAPK signaling pathway, p38 MAPK is particularly responsive to stress stimuli and plays critical roles in inflammatory responses, apoptosis, and differentiation. It is often activated by cellular stressors, influencing multiple downstream signaling events.
- **Jnk (c-Jun N-terminal kinase)**: A member of the MAPK family that responds to a variety of stress stimuli, including cytokines and environmental stress. JNK is involved in regulating apoptosis and the inflammatory response and can influence gene expression by activating specific transcription factors.

Each of these variables plays a vital role in cellular signaling networks and contributes to the complex interplay of biological processes that govern cell behavior in response to external stimuli. Understanding their specific functions and interactions is essential for unraveling the intricate web of signaling pathways critical for maintaining cellular homeostasis and responding to environmental changes.

2.2 Possible Causal Relations among these Variables

- **Raf** → **Mek**: Raf activates Mek through phosphorylation, initiating the MAPK signaling cascade that regulates cellular processes.
- **Mek** → **Erk**: Mek activates Erk by phosphorylating it, thus propagating the MAPK pathway for signal transduction.
- **Plcg** → **PIP2** → **PIP3**: Phospholipase C gamma hydrolyzes PIP2 to generate signaling molecules, producing PIP3 that acts as a secondary messenger.
- **PIP3** → **Akt**: PIP3 recruits and activates Akt at the plasma membrane, promoting cell survival and growth signaling.
- **DAG** → **PKC**: Diacylglycerol activates Protein Kinase C, which is involved in various signaling cascades related to growth and differentiation.
- **Akt** → **PKA**: Akt can phosphorylate and activate PKA, linking the survival and growth pathways to cAMP-mediated signaling.
- **Akt** → **Erk**: Akt signaling can influence Erk activation, creating potential cross-talk between the PI3K and MAPK pathways.
- **Jnk** → **Erk**: Jnk can modulate Erk activity under stress conditions, integrating responses to various stimuli.
- **P38** → **Mek and Erk**: P38 can interact with Mek and Erk during stress responses, affecting the MAPK pathway dynamics.
- **Erk** ↔ **Jnk and P38**: Erk, Jnk, and P38 participate in a network where activation of one may impact the others, highlighting intricate regulatory feedback mechanisms.
- **Plcg** → **PIP2**: Upon receptor activation, Plcg converts PIP2 into secondary messengers that activate multiple downstream pathways, reinforcing its role as a signaling hub.
- **PKC** → **Erk and P38**: PKC activates Erk and P38, further integrating diverse signaling responses across different cellular contexts.

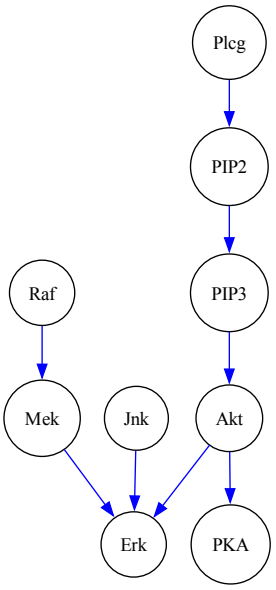


Figure 1: Possible Causal Relation Graph

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 Ramseys 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.

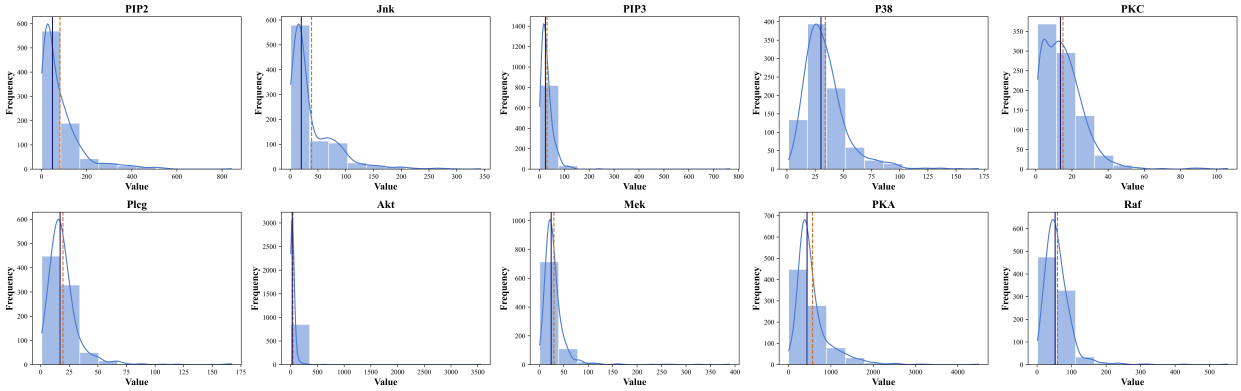


Figure 2: Distribution Plots of Variables

- 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 ($r > 0.8$), Moderate correlations ($0.5 < r < 0.8$), and Weak correlations ($r < 0.5$).

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

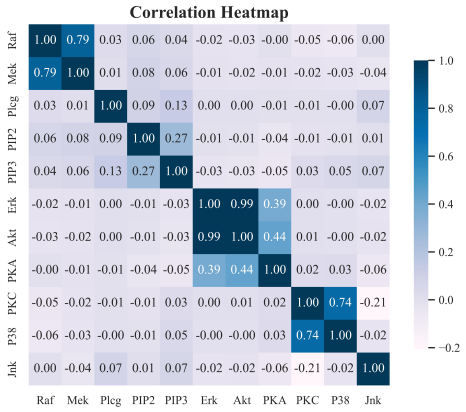


Figure 3: Correlation Heatmap of Variables

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**: Given the dataset's large sample size of 853 and the absence of missing values, the PC algorithm is suitable. It is efficient for large datasets and operates under the assumption of causal sufficiency, making it a good option when all relevant variables are observed.
- **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 ideal for this dataset due to its efficiency in dealing with continuous, non-Gaussian data. The sample size allows for effective score-based optimization while the non-linear relationships do not deter its performance significantly.
- **FCI**:
  - **Description**: The FCI algorithm is an extension of the PC algorithm designed to handle the presence of hidden confounders and outputs a Partial Ancestral Graph (PAG).
  - **Justification**: Despite the dataset not being explicitly heterogeneous, FCI is recommended as it allows for the possibility of hidden confounders. This is pertinent given the complexity and interconnectedness of biological pathways represented in the variables, leading to potentially unmeasured influences.

4.3 Hyperparameter Values Proposal assisted with LLM

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

- **alpha**:
  - **Value**: 0.1
  - **Explanation**: Given that the sample size is 853, which is relatively large, increasing the significance level to 0.1 allows for the detection of true edges without being overly conservative, thus reducing the chance of missing significant relationships due to the large sample context.
- **indep_test**:
  - **Value**: fisherz
  - **Explanation**: The Fisher's Z test is appropriate for continuous data, which aligns with the data type of the dataset. Although the relationships are not predominantly linear and Gaussian errors are not present, using Fisher's method could still provide a useful starting point for examining independence between variables.
- **uc_rule**:
  - **Value**: 0
  - **Explanation**: Using 0 for the unshielded colliders is suitable for the standard PC approach. Since the dataset does not show heterogeneity, this setting offers a basic structure without additional complexity that may not be needed.
- **uc_priority**:

  - **Value**: 2
  - **Explanation**: Prioritizing stronger colliders with a value of 2 strikes a balance between conservativeness and adaptability in determining relationships when conflicts arise, ensuring a structured approach given the dataset characteristics.

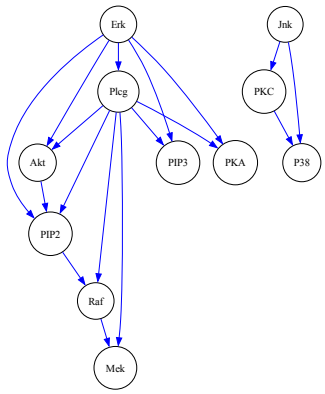
4.4 Graph Tuning with LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the LLM. We utilize LLM to help us determine the direction of undirected edges according to its knowledge repository. By integrating insights from the LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

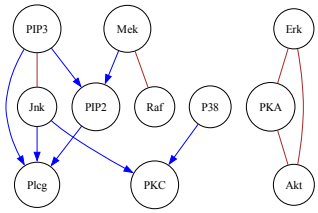
- **Raf** $\rightarrow$ **Mek**: Raf activates Mek through phosphorylation.
- **Jnk** $\rightarrow$ **PIP3**: Jnk can be activated by various stress stimuli, which includes the signaling pathway influenced by PIP3.
- **PKA** $\rightarrow$ **Erk**: Erk can influence PKA signaling pathways, but appears to be downstream in the context of signaling interactions.
- **Akt** $\rightarrow$ **PKA**: Akt can have downstream effects that activate PKA signaling.
- **Akt** $\rightarrow$ **Erk**: Erk can influence Akt activation through its signaling pathways, indicating a potential upstream role depending on context.

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

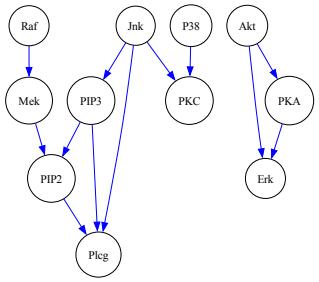
5 Results Summary



(a) True Graph



(b) Initial Graph



(c) Revised Graph

Figure 4: Graphs Comparison of PC

The above are result graphs produced by our algorithm. The initial graph is the graph in the first attempt, and the revised graph is the one pruned with LLM suggestion.

The causal relationships among the variables reveal a complex interdependence within the signaling pathways. Mek appears to exert a dual influence, both activating Raf and being regulated by it, indicating a feedback loop that is crucial for cellular signaling dynamics. Additionally, Mek drives the activation of Erk, a key player in the MAPK pathway, which further impacts the activity of both Akt and PKA, two important mediators of cellular survival and metabolism. PIP3, a critical phospholipid, is shown to be influenced by Mek and is also a signaling hub that affects Plcg, Jnk, and PKC, highlighting its role in various downstream signaling cascades. The intricate interactions extend to P38, which influences PKC as well, while Jnk also regulates multiple pathways by impacting Plcg and PIP3. Together, these relationships form a network that underscores the complexity of cellular signal transduction and the functional interplay of these vital components.

5.1 Graph Reliability Analysis

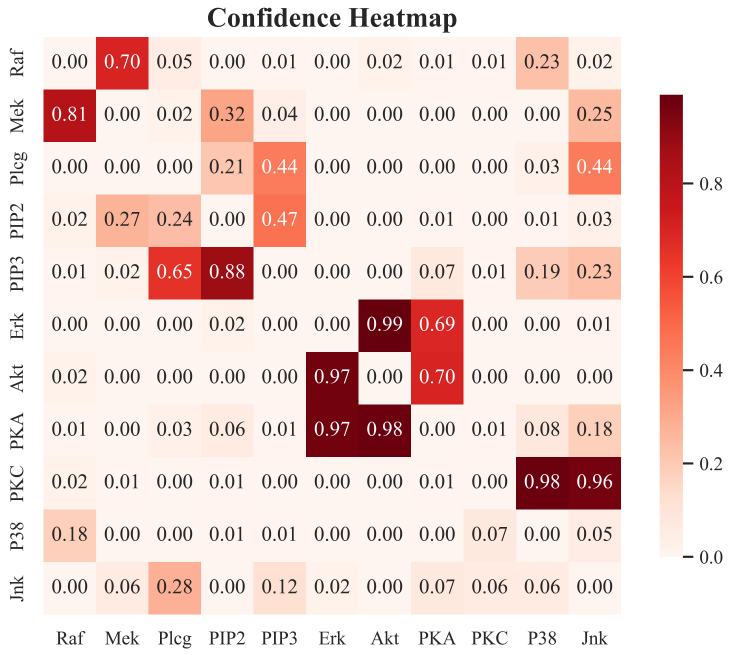


Figure 5: Reliability Graph

Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

From the statistics perspective, we have high confidence to believe that these edges exist: Erk $\rightarrow$ Akt (0.99), Akt $\rightarrow$ Erk (0.97), PKA $\rightarrow$ Akt (0.98), and PIP3 $\rightarrow$ PIP2 (0.88). Additionally, edges like Raf $\rightarrow$ Mek (0.7) and Mek $\rightarrow$ Erk (0.0) show varying confidence, with the latter suggesting no causal relationship, while others such as Mek $\rightarrow$ PIP2 (0.32) and Jnk $\rightarrow$ PKC (0.06) indicate low confidence in existence. Notably, edges with exceedingly low bootstrap probabilities, such as Mek $\rightarrow$ Erk (0.0) and PIP3 $\rightarrow$ Akt (0.0), support the conclusion that these relationships likely do not exist.

However, based on expert knowledge, we know that essential causal relationships, such as Raf $\rightarrow$ Mek, are well documented in the MAPK pathway, reinforcing its existence despite the moderate statistical confidence of 0.7. Likewise, the activation of Erk by Mek is established, making the reported 0

Therefore, the result of this causal graph is not entirely reliable. While some edges possess robust statistical backing, contradictions arise with established biological understandings, indicating that additional experimental validation and contextual analysis are necessary to enhance the fidelity of the causal relationships presented in this graph.