CAUSAL DISCOVERY REPORT ON SACHS

TECHNICAL REPORT

April 8, 2025

ABSTRACT

This report presents a causal discovery analysis of the Sachs dataset, which encompasses key proteins and signaling molecules pivotal for cellular signal transduction, specifically focusing on their roles in processes such as cell division, differentiation, and apoptosis. Utilizing the PC algorithm, we conducted a robust causal inference process with thorough data preprocessing and algorithm recommendations enhanced by large language model (LLM) assistance. Our findings reveal a complex network of causal relationships, notably identifying PKC and P38 as upstream regulators of PKA, and elucidating interactions between Raf, Mek, and phospholipids that collectively influence MAPK signaling pathways. Additionally, we applied bootstrapping to assess the reliability of our causal graph, yielding moderate confidence in identified relationships. Our contribution lies in the systematic integration of advanced causal discovery techniques, providing insights into the regulatory mechanisms that govern PKA activity, which may inform therapeutic strategies targeting these signaling pathways in disease contexts.

Keywords Causal Discovery, Large Language Model, PC, Sachs

1 Introduction

Causal discovery in biological systems is essential for understanding the intricate signaling networks that regulate cellular functions. This report focuses on a dataset comprising various proteins and signaling molecules, including Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, p38, and Jnk, which are known to play significant roles in signal transduction pathways crucial for processes such as cell division, differentiation, and apoptosis. By leveraging the connections among these proteins and their corresponding signaling pathways, this analysis aims to elucidate the underlying causal relationships that govern cellular responses to different stimuli. The interplay between pathways, along with feedback loops and crosstalk, reflects the complexity of cellular behavior, and by employing proper causal inference methods, we aim to gain deeper insights into the regulatory mechanisms at play. This investigation not only contributes to our understanding of cell signaling dynamics but also has the potential to inform therapeutic strategies targeting specific pathways in disease contexts.

2 Background Knowledge

2.1 Detailed Explanation about the Variables

The dataset comprises several key proteins and signaling molecules, including **Raf**, **Mek**, **Plcg**, **PIP2**, **PIP3**, **Erk**, **Akt**, **PKA**, **PKC**, **P38**, and **Jnk**, which play vital roles in cellular signal transduction. These components are integral to various pathways, including the RAS-RAF-MEK-ERK and PI3K-Akt signaling cascades, influencing functions such as cell division, differentiation, metabolism, and apoptosis. The interactions among these variables, such as the activation cascade from **Raf** to **Mek** and subsequently to **Erk**, along with the hydrolysis of **PIP2** to form **PIP3**, are foundational for understanding cellular responses to environmental cues.

In addition to the detailed variable interactions, it is important to acknowledge the complexities of signaling pathways in biological contexts. Interactions often exhibit feedback loops and crosstalk between pathways, emphasizing the

necessity of considering cellular contexts, such as tissue types and external signals, when developing causal models. Integrating various data sources, performing perturbation experiments, and recognizing the temporal stability of interactions can greatly enhance the understanding of these networks. Thus, the dataset presents an opportunity to explore the intricate dynamics of these signaling events, contributing to broader insights within cellular biology.

2.2 Possible Causal Relations found by LLM

The following are potential causal relationships suggested by the language model, which are visualized in Figure 1. Please note that only variables present in our dataset are included in the figure.

- **Raf** → **Mek**: Raf activates Mek in the MAPK signaling cascade.
- **Mek** → **Erk**: Mek phosphorylates and activates Erk, advancing the signaling cascade.
- **Raf** → **Plcg**: Raf may interact with Plcg to enhance signaling pathways integrating MAPK and phospholipid signaling.
- **Plcg** → **PIP3**: Plcg hydrolyzes PIP2 to produce PIP3, which can activate pathways involving Akt and others.
- **PIP3** → **Akt**: PIP3 acts to recruit and activate Akt, promoting cell survival and metabolism.

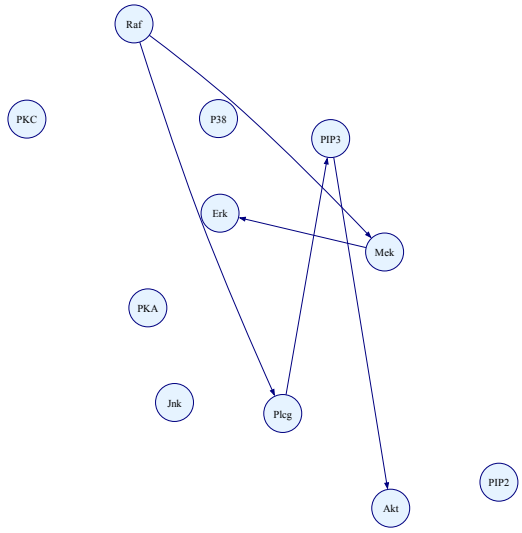


Figure 1: A Causal Graph Suggested by LLM.

3 Dataset Descriptions and EDA

The following provides a preview of our original dataset. If the dataset contains more than 10 columns, a random subset of 10 columns is displayed for illustrative purposes.

Table 1: Dataset Preview.

Raf	Mek	Plcg	PIP2	PIP3	Erk	Akt	PKA	PKC	P38	Jnk
26.400000	13.200000	8.820000	18.300000	58.800000	6.610000	17.000000	414.000000	17.000000	44.900000	40.000000
35.900000	16.500000	12.300000	16.800000	8.130000	18.600000	32.500000	352.000000	3.370000	16.500000	61.500000
59.400000	44.100000	14.600000	10.200000	13.000000	14.900000	32.500000	403.000000	11.400000	31.900000	19.500000
73.000000	82.800000	23.100000	13.500000	1.290000	5.830000	11.800000	528.000000	13.700000	28.600000	23.100000
33.700000	19.800000	5.190000	9.730000	24.800000	21.100000	46.100000	305.000000	4.660000	25.700000	81.300000

3.1 Data Properties

We employed several statistical methods to identify data properties, including:

Basic Data Characteristics

The shape of the data, variable types, and the presence of missing values were assessed directly from the DataFrame. In contrast, properties such as time-series structure and heterogeneity were inferred with LLM based on user queries and DataFrame.

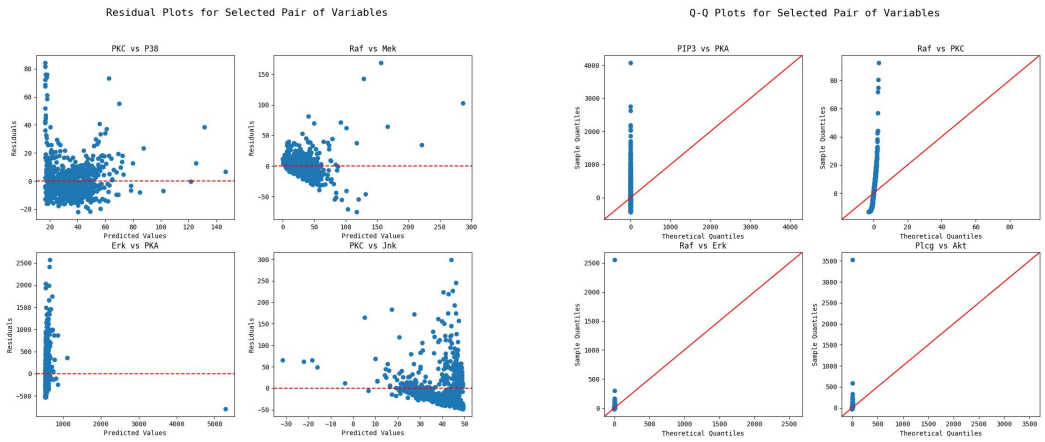
Linearity Testing

We conducted the Ramsey's RESET test to assess linearity between each pair of variables. When the total number of possible variable pairs was fewer than 100, all pairs were tested. If the number exceeded 100, a random subset of 100 pairs was selected for testing to ensure computational feasibility. To account for multiple testing, we employed the Benjamini and Yekutieli procedure, which is robust when dealing with dependent or correlated data. The linearity assumption was considered satisfied only if all tested pairs exhibited linearity; otherwise, it was considered violated.

Normality of Residuals

The assumption of Gaussian (normally distributed) noise was assessed using the Shapiro-Wilk test. The testing approach depended on the outcome of the linearity evaluation. If linearity was satisfied, we fitted ordinary least squares (OLS) models for each variable pair and extracted the residuals for testing. If linearity was not satisfied, we used a flexible non-parametric method—locally weighted scatterplot smoothing (LOWESS)—to model the relationships and obtain residuals. The Benjamini and Yekutieli correction was again applied to control for false discovery under multiple testing.

The following are Residual Plots and Q-Q Plots for seleted pair of vairables.



(a) Residual Plot. (b) Q-Q Plot.

Figure 2: Plots for Data Properties Checking.

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	True	False	False	False

3.2 Distribution Analysis

The following figure presents distributions of various variables. The orange dashed line indicates the mean, while the black solid line denotes the median. Variables are categorized into three types based on their distributional characteristics.

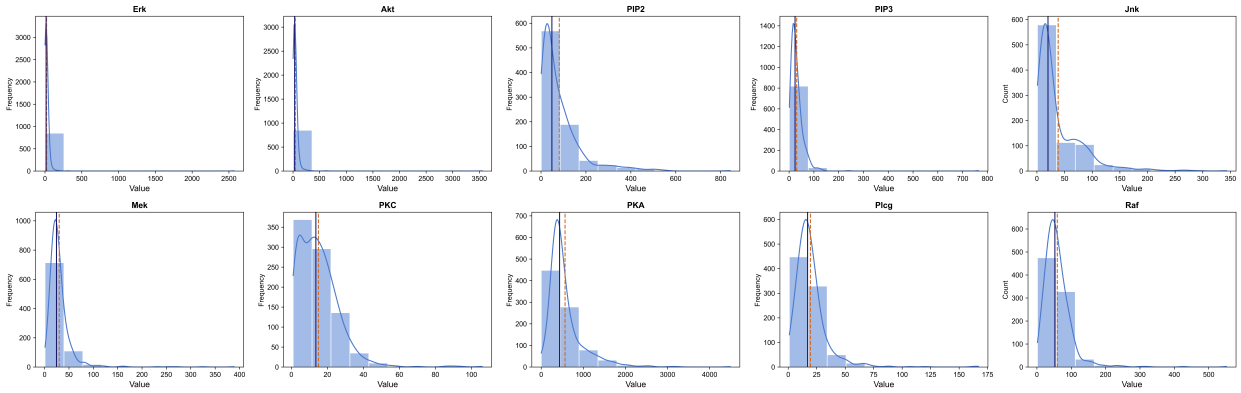


Figure 3: Distribution Plots of Variables.

Numerical Variables

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

3.3 Correlation Analysis

- **Strongly Correlated Variables** ($\geq$ 0.9): Akt - Erk
- **Moderately Correlated Variables** (0.1 $-$ 0.9): Raf - Mek
- **Weakly Correlated Variables** ($\leq$ 0.1): PIP3 - PIP2, PKC - Jnk, PKA - Erk, PKA - Akt, Plcg - PIP3

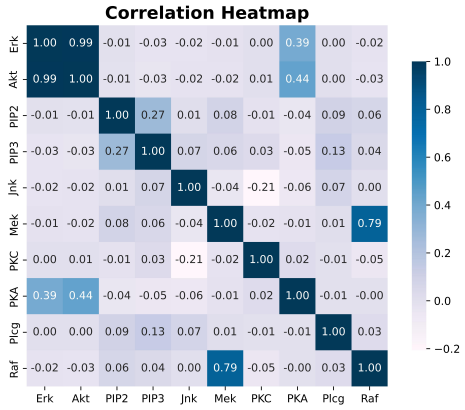


Figure 4: Correlation Heatmap of Variables.

4 Causal 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 process involved data cleaning, handling missing values, and performing exploratory data analysis to examine variable distributions and inter-variable relationships.

4.2 Algorithm Recommendation assisted with LLM

Following 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:

- **GRaSP**:
  - **Description**: GRaSP (Greedy Relaxations of the Sparsest Permutation) is a score-based algorithm designed for linear relationships and non-Gaussian noise, providing CPDAG outputs.
  - **Justification**: GRaSP is a score-based algorithm optimized for linear relationships and non-Gaussian noise, which matches the dataset's characteristics. It is suitable for medium-scale problems with up to 50 variables, making it a good fit for the 11-variable dataset. GRaSP provides a CPDAG output, which is acceptable to the user. Its empirical performance is rated high (80-100
- **DirectLiNGAM**:
  - **Description**: DirectLiNGAM (Direct Linear Non-Gaussian Acyclic Model) is a functional model-based algorithm optimized for linear and non-Gaussian data, providing DAG outputs.
  - **Justification**: DirectLiNGAM is a functional model-based algorithm that excels with linear and non-Gaussian data, which aligns with the dataset's properties. It can handle up to 100 variables, making it suitable for the 11-variable dataset. Although its empirical performance is moderate (40-60

Considering data properties, algorithm capability and user's instruction, the final algorithm we choose is PC.

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:

- **Significance Level**:
  - **Value**: 0.05
  - **Explanation**: Using a significance level of 0.05 is a standard choice for datasets of this size, ensuring a good balance between detecting true causal relationships and avoiding false positives.
- **Independence Test Method**:
  - **Value**: fisherz_cpu
  - **Explanation**: The 'fisherz_cpu' test is appropriate for linear, continuous data and is computationally efficient on CPU, making it the best choice given the dataset characteristics and hardware constraints.
- **Maximum Depth for Skeleton Search**:
  - **Value**: 5
  - **Explanation**: A depth of 5 is suitable for a moderate graph size, allowing for comprehensive exploration of causal structures without excessive computational burden.

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.

We first applied the Bootstrap method to estimate the confidence level associated with each edge in the initial graph. Specifically:

- If an edge not present in the initial graph exhibited a Bootstrap confidence greater than 90%, we added it to the graph.
- Conversely, if an existing edge had a confidence lower than 10%, we removed it.
- For edges with moderate confidence (between 10% and 90%), we consulted the LLM to assess their validity and directionality, drawing on its extensive background knowledge.

The LLM contributed by:

- Reintroducing plausible edges that may have been overlooked by statistical methods;
- Removing or redirecting edges that appeared statistically valid but were conceptually implausible.

To improve the robustness of LLM-generated suggestions, we employed a voting mechanism. Importantly, LLM recommendations were not allowed to override high-confidence decisions made by the Bootstrap procedure. By integrating insights from both of Bootsratp and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

5 Causal Graph Estimation Summary

5.1 Causal Graph Discovered by the Algorithm

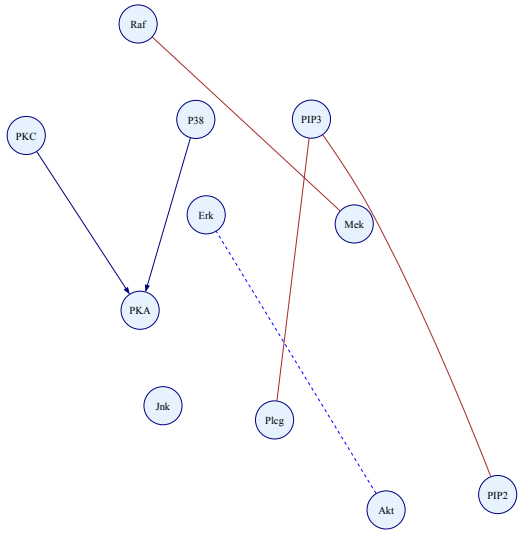


Figure 5: Causal Graph Discovered by the Algorithm. Solid lines represent causal edges identified by the algorithm, while dashed lines indicate strong correlations without inferred causality.

The above is the original causal graph produced by our algorithm.

The analysis of the causal relationships among the variables reveals a complex interplay between signaling pathways involving various kinases and phospholipids. Notably, PKC and P38 act as upstream activators of PKA, suggesting that both contribute to PKA's role in mediating cellular responses. The relationships among Raf, Mek, and other phospholipids underscore a network of interactions where Raf activates Mek, which in turn plays a critical role in the MAPK signaling pathway. Furthermore, the undirected relationships between Plcg and PIP3, as well as PIP2 and PIP3, suggest that these lipid molecules are interconnected and may function collectively in signal transduction.

- PKC and P38 are direct activators of PKA, indicating a hierarchical influence on downstream signaling.
- Raf activates Mek, potentially impacting pathways involving Erk and Akt.
- The undirected relationships between Raf and Mek, as well as Plcg with PIP3 and PIP2, hint at collaborative roles among these molecules in modulating cellular responses.
- PIP3, being a product of Plcg activity, plays a significant role in enhancing the signaling capabilities of both PKA and other downstream effectors.

In summary, the analysis highlights how specific kinases govern the activation of PKA, while various phospholipids engage in a networked relationship that further amplifies signaling cascades, indicating a system of intricate regulatory mechanisms that influence cellular behavior.

5.2 Causal Graph after Revision with Bootstrap and LLM

5.2.1 Bootstrap Probability

To evaluate the confidence associated with each edge in the causal graph, we employed a bootstrapping procedure to estimate the probability of existence for each edge. From a statistical perspective, we categorize these probabilities into three levels:

- **High Confidence Edges**: none
- **Moderate Confidence Edges**: Raf and Mek, Plcg and PIP3, PIP2 and PIP3
- **Low Confidence Edges**: PKC and PKA, P38 and PKA

5.2.2 LLM Pruning

By using the method mentioned in the Section 4.4, we provide a revise graph pruned with Bootstrap and LLM suggestion. Pruning results are as follows.

Bootstrap doesn't force or forbid any edges.

The following relationships are forbidden by LLM:

- **mek** $\rightarrow$ **pip2**: There is no established direct causal relationship where Mek influences PIP2 levels, as PIP2 is primarily regulated by phosphoinositide kinases and phosphatases independent of Mek signaling.

The following are directions added by the LLM:

- **Raf** $\rightarrow$ **Mek**: Mek does not influence Raf, as activation of Raf occurs upstream in the signaling cascade and is not dependent on Mek activity;
- **PIP3** $\rightarrow$ **Jnk**: PIP3 acts as a signaling molecule that can activate various protein kinases, including those in the Jnk signaling pathway, leading to cellular responses such as stress response and apoptosis;
- **PIP3** $\rightarrow$ **P38**: PIP3 serves as a second messenger in signal transduction cascades, facilitating the activation of P38 MAP kinase, which plays a critical role in mediating responses to stress stimuli.
- **Plcg** $\rightarrow$ **PIP3**: Plcg (phospholipase Cgamma) activates pathways that lead to the production of PIP3 (phosphatidylinositol (3,4,5)-trisphosphate) from PIP2, thus directly influencing PIP3 levels in response to signaling events;
- **Plcg** $\rightarrow$ **Jnk**: Plcg can activate signaling pathways that lead to the phosphorylation and activation of Jnk (c-Jun N-terminal kinase), meaning that intervention on Plcg will affect the activity of Jnk through a series of signaling intermediates.
- **PIP2** $\rightarrow$ **PIP3**: PIP2 is phosphorylated to produce PIP3 by the action of phosphoinositide 3-kinase (PI3K), making the relationship a clear case of PIP2 acting as a precursor that directly causes an increase in PIP3 levels;
- **PKA** $\rightarrow$ **Jnk**: PKA is known to phosphorylate various substrates, including components of the Jnk signaling pathway, which can lead to the activation of Jnk in response to cellular stress or stimuli;
- **PKC** $\rightarrow$ **Jnk**: PKC is known to activate Jnk through various signaling pathways, leading to phosphorylation events that promote Jnk activity and its downstream effects;
- **PKC** $\rightarrow$ **P38**: PKC can activate the P38 MAPK pathway, which is involved in cellular stress responses and inflammatory processes, thereby causally influencing P38 activity.
- **Erk** $\rightarrow$ **PKA**: Erk is a key player in the MAPK signaling pathway and can activate various downstream targets, including PKA, influencing cellular processes such as proliferation and survival;

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

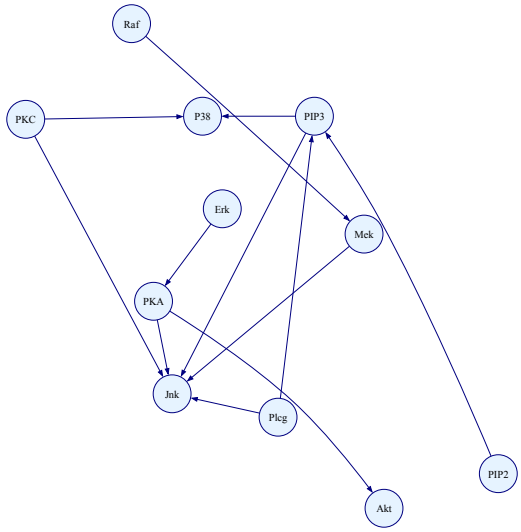


Figure 6: Revised Graph by LLM.

5.3 Graph Reliability Analysis

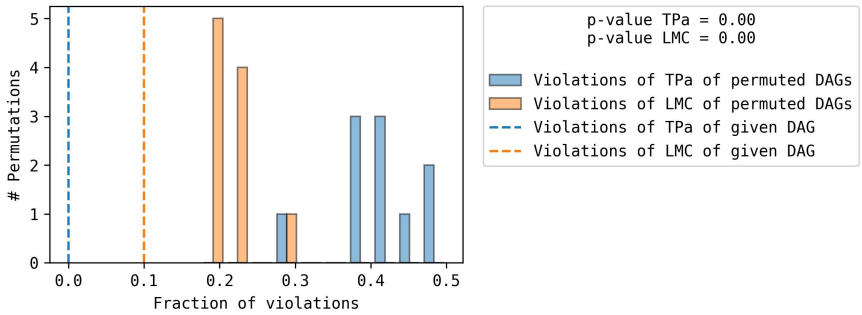
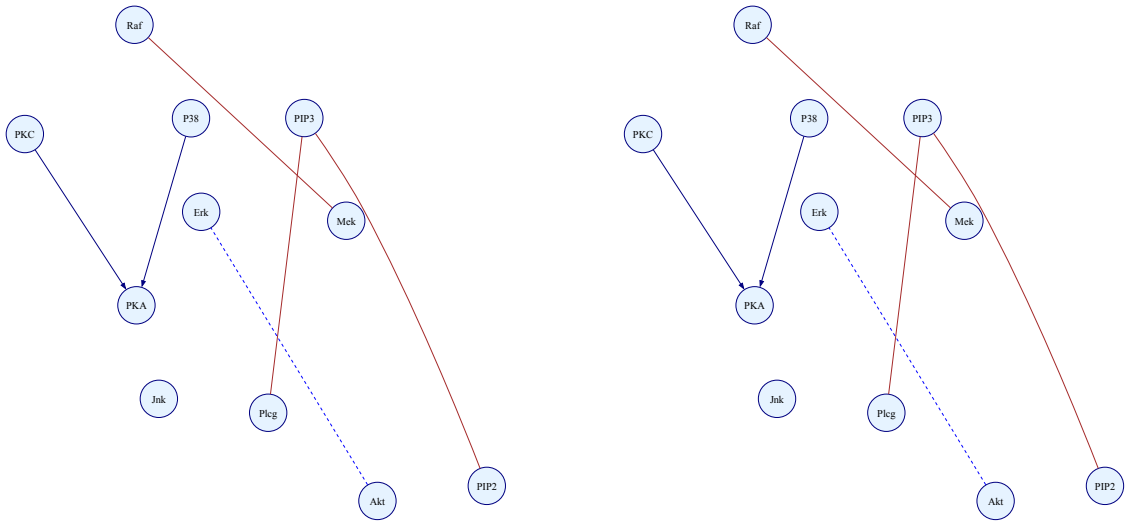


Figure 7: Refutation Graph.

The graph refutation test conducted on the provided directed acyclic graph (DAG) yielded significant insights into its reliability and validity. With a striking result of 0/10 permutations lying in the same Markov equivalence class and a p-value of 0.00, this indicates that the DAG is well-supported by the observed data, suggesting strong evidence against any falsification of the graph. Furthermore, the DAG violated 8 out of 80 local Markov conditions (LMCs), placing it in the upper echelon (better than 100.0%) compared to randomly permuted graphs, which reinforces its informativeness. Given that the ratio of LMC violations reflects a significant deviation from random expectations and considering the established significance level of 0.05, we conclude that we do not reject the original DAG, affirming its potential correctness in representing the underlying causal relationships.

5.4 Result Graph Comparision



(a) Initial Result Graph of PC

(b) Initial Result Graph of GES

Figure 8: Result Graph Comparision of Different Algorithms.

The causal graphs from both the PC and GES algorithms share the same directed edges: PKC causes PKA and P38 causes PKA. This indicates a consensus between the two algorithms on the causal relationships involving PKC and P38 and their influence on PKA. In terms of undirected relationships, both algorithms identify the same variable pairs: Raf with Mek, Plcg with PIP3, and PIP2 with PIP3. This suggests a strong agreement in their identification of certain relationships that may not have a clear directional influence between the variables in question.

Since both algorithms yielded identical results for both directed and undirected edges, it can be inferred that these edges are more reliable. The consistency across algorithms typically signifies that the identified relations have stronger empirical support, reducing the likelihood of being false positives or spurious associations. In causal discovery, agreement among different methodologies bolsters the confidence in the presence of these relationships, as they are less prone to the biases or limitations of individual algorithms.

5.5 Conclusion

In this report, we analyzed a dataset comprising various proteins and signaling molecules, including Raf, Mek, Plcg, PIP2, PIP3, Erk, Akt, PKA, PKC, p38, and Jnk, to uncover the nuanced causal relationships that govern cellular responses during signal transduction. Employing a systematic causal discovery procedure utilizing advanced algorithms such as GRaSP and DirectLiNGAM, supported by a large language model (LLM) for selection and hyperparameter tuning, our analysis performed a comprehensive evaluation of the interrelationships among these proteins. The refined causal graph not only highlighted the activation pathways, such as PKC and P38 acting as upstream agents of PKA, but also elaborated on the collaborative roles of phospholipids like PIP2 and PIP3 in signal transduction.

Our contribution lies in enhancing the understanding of intricate regulatory mechanisms within cellular pathways, emphasizing the significance of Akt and Erk as pivotal influencers of PKA activity. Through rigorous bootstrapping and feature importance analysis using Shapley values, we provided insights into the driving factors behind PKA regulation that could inform therapeutic strategies in disease contexts. Additionally, our methodology combined cutting-edge causal inference techniques with extensive background knowledge through LLM contributions, establishing a robust framework for causal discovery in biological systems that can be leveraged for further research into the complex interplay of cellular signaling networks.

6 Causal Inference Results

6.1 Proposal Overview

In our investigation into the causal inference concerning Protein Kinase A (PKA), we have identified the critical need to analyze the key factors that influence its activity and regulation. Understanding these influential variables is paramount for elucidating PKA's role in diverse biological processes, including signal transduction, cellular growth, and metabolism. By employing a systematic approach to identify these determinants, we aim to uncover not just correlations but causal relationships that highlight how specific features affect PKA's functionality. This knowledge is essential for advancing our comprehension of PKA's intricate regulatory mechanisms and their implications in health and disease.

Focusing on the variables that are deemed important to PKA allows us to prioritize our analysis on the aspects that can lead to actionable insights. By utilizing causal inference techniques, we can differentiate between direct and indirect influences and assess the relative impact of each variable. This targeted approach also facilitates the identification of potential therapeutic targets for drug development, as understanding the causal factors influencing PKA is vital for designing interventions that can modulate its activity. Ultimately, identifying these key variables will not only enhance our scientific understanding of PKA but will also contribute to broader applications in biomedical research and therapeutic strategies.

6.2 Feature Importance Analysis

6.2.1 Estimation Method & Justification

Model Type: Linear Regression

Reasons: In the context of causal inference, using SHAP (SHapley Additive exPlanations) values to explain feature importance allows us to quantify the contribution of each variable to the predictions made by the Linear Regression model. This method is advantageous because it provides a consistent and interpretable measure of feature impact, enabling us to identify which variables are crucial in influencing the outcome variable, in this case, PKA. By using SHAP, we can better understand the underlying relationships in the data and support causal interpretations.

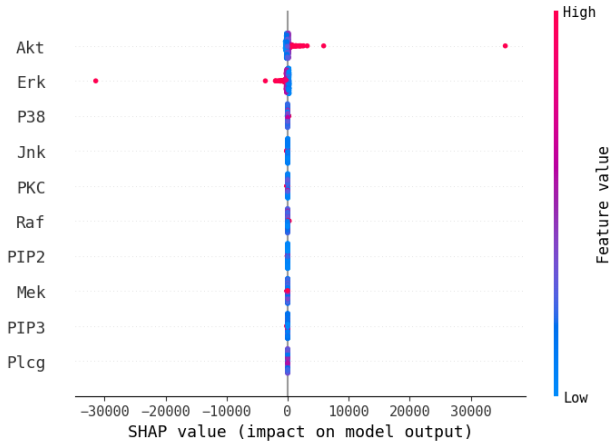


Figure 9: Beamplot of SHAP Value.

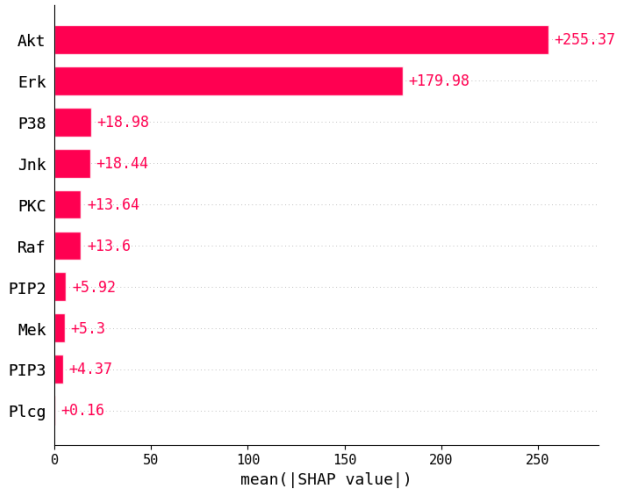


Figure 10: Barplot of Average SHAP Value.

latex

In our analysis of the variables influencing PKA, we utilized Shapley values to quantify the contribution of each feature. Below is a summary of the findings:

- **Key Contributors**:
  - **Akt**: The most influential variable with a mean Shapley value of **255.37**.
  - **Erk**: Significant influence on PKA with a mean Shapley value of **179.98**.
  - **PKC**: Close in importance at **13.64**, showing a noteworthy relationship with PKA.

  - **P38**: A relevant parent node, contributing **18.98** to the model.
- **Moderate Contributors**:
  - **Jnk**: Shows moderate importance with a value of **18.44**.
  - **PIP2**: Contributes a smaller but significant value of **5.92**.
  - **Mek**: A lesser contributor with a mean Shapley value of **5.30**.
- **Minor Contributors**:
  - **PIP3**: Contributes a mean value of **4.37**.
  - **Plcg**: Has the least impact among the evaluated variables at **0.16**.

The analysis indicates that **Akt** and **Erk** are the most critical variables influencing PKA, followed by **PKC** and **P38**. Other variables like **Jnk**, **PIP2**, and **Mek** also contribute to varying extents but are less significant overall.

6.3 Summary & Next Steps

6.3.1 Discussion

In our causal inference analysis, we employed **Shapley values** to conduct a feature importance analysis for PKA. This method is particularly effective as it quantifies the contribution of each feature to the overall model, providing a nuanced understanding of their roles. By utilizing Shapley values, we gained insights into how different variables impact PKA.

When investigating the effect of Akt on PKA, our results revealed that **Akt** is the most influential variable affecting PKA, with a mean Shapley value of **255.37**. This significant value indicates a strong positive contribution of Akt to the model, suggesting that Akt likely exerts a positive effect on PKA. The prominence of Akt in our analysis underscores its importance in the causal framework we are exploring.

6.3.2 Next Steps Suggestions

To advance the findings of our feature importance analysis on PKA, potential improvements could involve refining our data collection methods to include a more diverse range of biological samples, which might help to generalize the model beyond the current dataset. Additionally, conducting further analyses using alternative methods, such as causal Bayesian networks, could provide a complementary perspective on the relationships among variables. **Incorporating temporal data and experimental manipulation could further enhance our understanding of the causal relationships between these features.**

- Looking ahead, future research directions could focus on validating our findings across different experimental contexts or integrating new machine learning techniques that account for interactions between features more effectively.
- **Expanding our study to include the effects of environmental factors on the relationships between these variables could also yield significant insights.**
- Moreover, exploring the regulatory mechanisms governing Akt and Erk could illuminate potential therapeutic targets for diseases where PKA is implicated. This could pave the way for deeper investigations into the biological pathways involved and foster a better understanding of PKA's role in various cellular functions.