

---

# CAUSAL DISCOVERY REPORT ON ABALONE

---

## TECHNICAL REPORT



April 8, 2025

## ABSTRACT

This report presents a thorough causal discovery analysis of a dataset encompassing various biological and morphological metrics related to abalones, specifically focusing on Age, Length, Shell weight, Diameter, Height, Whole weight, Shucked weight, and Viscera weight. We employed the Fast Causal Inference (FCI) algorithm following comprehensive data preprocessing and exploration, supported by a large language model (LLM) for algorithm selection and hyperparameter optimization. Our findings reveal that Age is a critical determinant influencing Length, Shell weight, Diameter, and Whole weight, with established correlations between Length and various weight metrics. The absence of D-separation among these variables suggests direct and indirect causal pathways, underscoring the complexity of these relationships. This analysis contributes significantly to understanding abalone growth dynamics, highlighting the necessity of considering Age and associated growth metrics in future ecological and biological studies. The integration of bootstrap techniques for edge confidence estimation further enhances the reliability of the inferred causal framework.

**Keywords** Causal Discovery, Large Language Model, FCI, Abalone

## 1 Introduction

The dataset under investigation offers valuable insights into abalones, an essential group of marine mollusks, and encompasses various biological and morphological variables critical to understanding their growth and health. Key attributes such as Age, Length, Shell\_weight, Diameter, Height, Whole\_weight, Shucked\_weight, and Viscera\_weight form a comprehensive framework to explore the relationships among these traits. Age serves as a vital predictor of size metrics, including Length, Diameter, and Height, while these dimensions, in turn, correlate with the overall body weight of the abalone. Additional insights can be gained by considering the potential effects of environmental factors, which, though not directly measured in the dataset, play a pivotal role in influencing growth and health. This report aims to conduct a thorough causal discovery and inference analysis informed by the intricacies of these relationships and providing a deeper understanding of the biological and ecological factors that drive the dynamics of abalone populations.

## 2 Background Knowledge

### 2.1 Detailed Explanation about the Variables

The dataset centers around various biological and ecological variables related to abalones, which are marine mollusks. Key variables include **Age**, estimated through physical measurements; **Length**, **Diameter**, **Height**, and **Shell weight**, all of which provide insights into the size, health, and growth of these creatures. Additionally, **Whole weight**, which encompasses the total body weight, and **Shucked weight** and **Viscera weight**, which focus on the abalone's meat and internal organs, respectively, are crucial for understanding both biological fitness and market value. These variables offer critical insights into the growth patterns and health conditions of abalones, as well as serve as vital indicators for the ecological sustainability and economic viability of abalone fishing.

To enhance causal discovery efforts, it is essential to consider background knowledge including **Growth Patterns**, which highlight variations in growth rates based on species and environmental conditions. Understanding **Market Valuation** contributes to recognizing which variables are most impactful to economic outcomes, particularly in market contexts. Furthermore, **Biological Interactions** and **Ecological Impacts**—incorporating factors like environmental stressors and predator-prey dynamics—are vital for contextualizing the biological data within broader ecological frameworks, thereby enriching the causal analysis of the dataset.

## 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.

- **Age → Length:** As abalones grow older, they typically increase in size, which means their length is likely to increase with age.
- **Age → Diameter:** As abalones grow older, they typically increase in size, which means their diameter is likely to increase with age.
- **Age → Height:** As abalones grow older, they typically increase in size, which means their height is likely to increase with age.

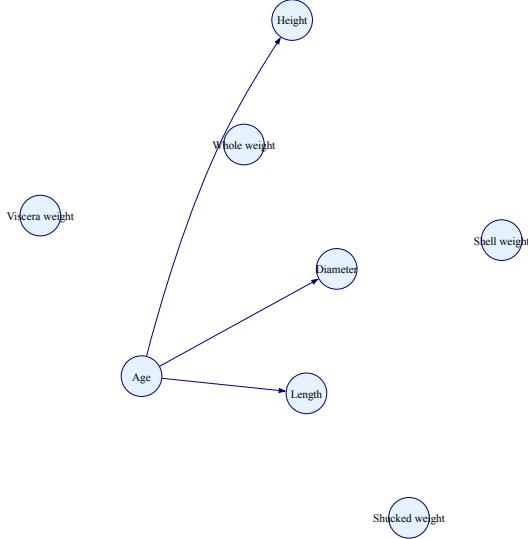


Figure 1: A Causal Graph Suggested by LLM.

## 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.

Age	Length	Shell weight	Diameter	Height	Whole weight	Shucked weight	Viscera weight
15.000000	0.455000	0.365000	0.095000	0.514000	0.224500	0.101000	0.150000
7.000000	0.350000	0.265000	0.090000	0.225500	0.099500	0.048500	0.070000
9.000000	0.530000	0.420000	0.135000	0.677000	0.256500	0.141500	0.210000
10.000000	0.440000	0.365000	0.125000	0.516000	0.215500	0.114000	0.155000
7.000000	0.330000	0.255000	0.080000	0.205000	0.089500	0.039500	0.055000

### 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 selected pair of variables.

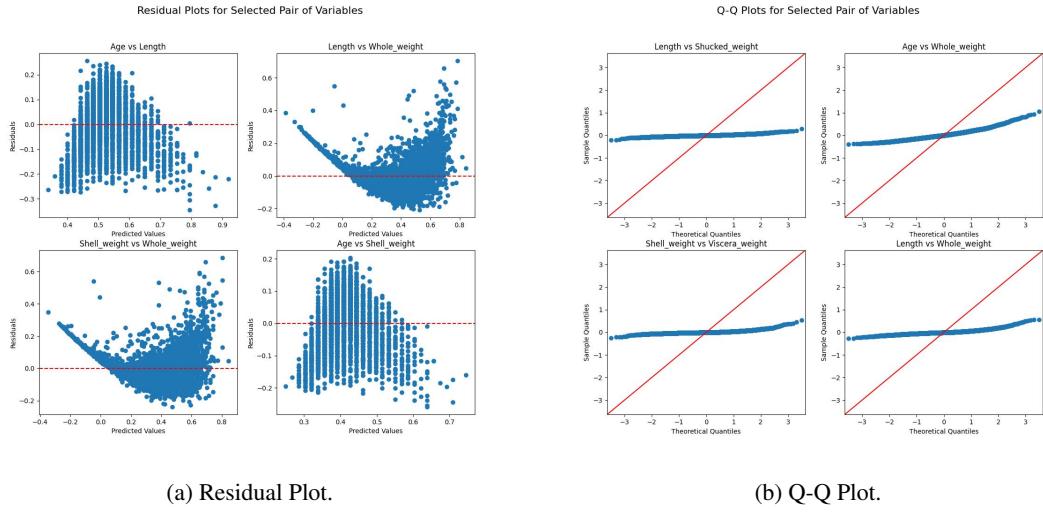


Figure 2: Plots for Data Properties Checking.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties.

Shape ( $n \times d$ )	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(4177, 8)	Continuous	False	False	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.

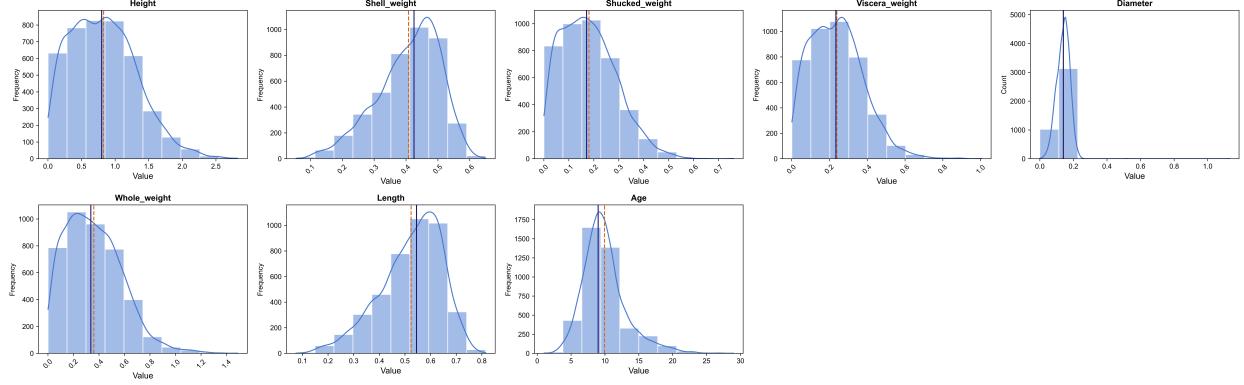


Figure 3: Distribution Plots of Variables.

### Numerical Variables

- Slight left skewed distributed variables: Shell weight, Diameter, Length
- Slight right skewed distributed variables: Height, Shucked weight, Viscera weight, Whole weight, Age
- Symmetric distributed variables: None

### 3.3 Correlation Analysis

- **Strongly Correlated Variables ( $\geq 0.9$ ):** Shell weight - Height, Shucked weight - Height, Shucked weight - Shell weight, Viscera weight - Height, Viscera weight - Shell weight, Viscera weight - Shucked weight, Diameter - Height, Diameter - Shell weight, Diameter - Viscera weight, Whole weight - Height, Whole weight - Shell weight, Whole weight - Shucked weight, Whole weight - Viscera weight, Length - Height, Length - Shell weight, Length - Shucked weight, Length - Viscera weight, Length - Diameter, Length - Whole weight, etc.
- **Moderately Correlated Variables (0.1 – 0.9):** Diameter - Shucked weight, Whole weight - Diameter, Age - Height, Age - Shell weight, Age - Shucked weight, Age - Viscera weight, Age - Diameter, Age - Length
- **Weakly Correlated Variables ( $\leq 0.1$ ):** Age - Whole weight

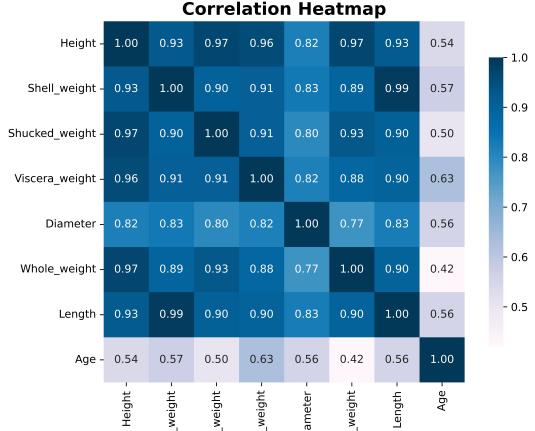


Figure 4: Correlation Heatmap of Variables.

## 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:

- **FCI:**
  - **Description:** Fast Causal Inference (FCI) is a constraint-based algorithm that can handle latent confounders and outputs a Partial Ancestral Graph (PAG), making it suitable for datasets with potential hidden variables.
  - **Justification:** The FCI algorithm is suitable because it is a constraint-based method that handles both linear and non-linear relationships effectively. It is robust to latent variables, which is beneficial if there are unmeasured confounders. FCI can output a PAG, which is acceptable to the user. It performs well with non-Gaussian noise and is efficient for the dataset's size.
- **IAMBnPC:**
  - **Description:** Incremental Association Markov Blanket with PC algorithm (IAMBnPC) is a hybrid approach that combines constraint-based and score-based methods, providing flexibility in handling various data characteristics and outputting a CPDAG.
  - **Justification:** The IAMBnPC algorithm is a hybrid method that is flexible in handling both linear and non-linear relationships. It is efficient for the dataset's size and outputs a CPDAG, which is acceptable to the user. It performs well with non-Gaussian noise and is robust in homogeneous data settings.

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

## 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 an alpha of 0.05 is a standard choice for moderate sample sizes, ensuring a balance between detecting true causal relationships and avoiding false positives.
- **Independence Test Method:**
  - **Value:** rcit
  - **Explanation:** The 'rcit' test is chosen for its ability to handle non-linear relationships efficiently, providing a good balance between accuracy and computational feasibility for the given dataset size.
- **Maximum Depth for Skeleton Search:**
  - **Value:** 6
  - **Explanation:** A depth of 6 allows for a comprehensive search of causal paths, maximizing accuracy by exploring deeper connections in the small graph.

## 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

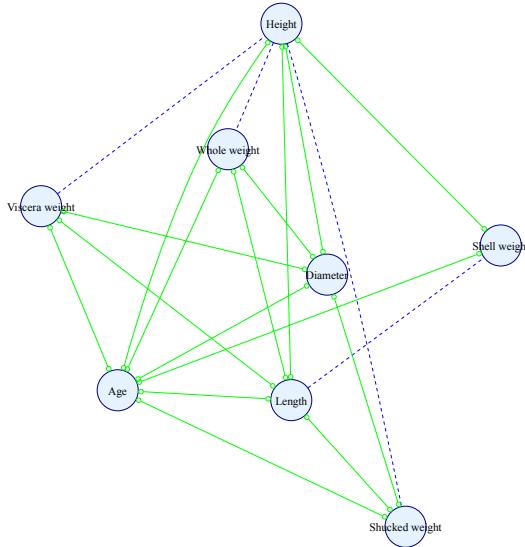


Figure 5: Causal Graph Discovered by the Algorithm. Solid lines represent causal edges identified by the algorithm, while dashed lines indicate strong correlations without inferred causality.

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

The analysis of the variable relationships reveals a complex interplay among the biological metrics often associated with marine organisms, particularly mollusks or shellfish. The findings suggest that Age is a central factor that significantly influences several other metrics, including Length, Shell weight, Diameter, and Whole weight. Length also appears to interact with various weight metrics, including Viscera weight, Shucked weight, and Shell weight, indicating a correlated growth pattern. Notably, Height seems to have close ties with Diameter and Length, further emphasizing its potential role in overall physical growth. The lack of D-separation among these variables implies direct or indirect causal pathways from Age to multiple attributes, reinforcing the understanding that as organisms age, their physical dimensions and corresponding weights inevitably change.

- Age strongly influences Length, Shell weight, Diameter, and Whole weight.
- Length correlates with various weights such as Viscera weight, Shucked weight, and Shell weight.
- Height is closely related to Diameter and Length, indicating its importance in growth metrics.
- The absence of D-separation indicates potential direct or indirect causal relationships among these variables.

In summary, the root cause of changes in Length and various weights appears to be the Age of the organism, which is further mediated by the growth metrics like Height and Diameter, demonstrating that Age is a critical variable in understanding the physical development and size characteristics of the species in question.

## 5.2 Causal Graph after Revision with Bootstrap and LLM

### 5.2.1 Bootstrap Probability

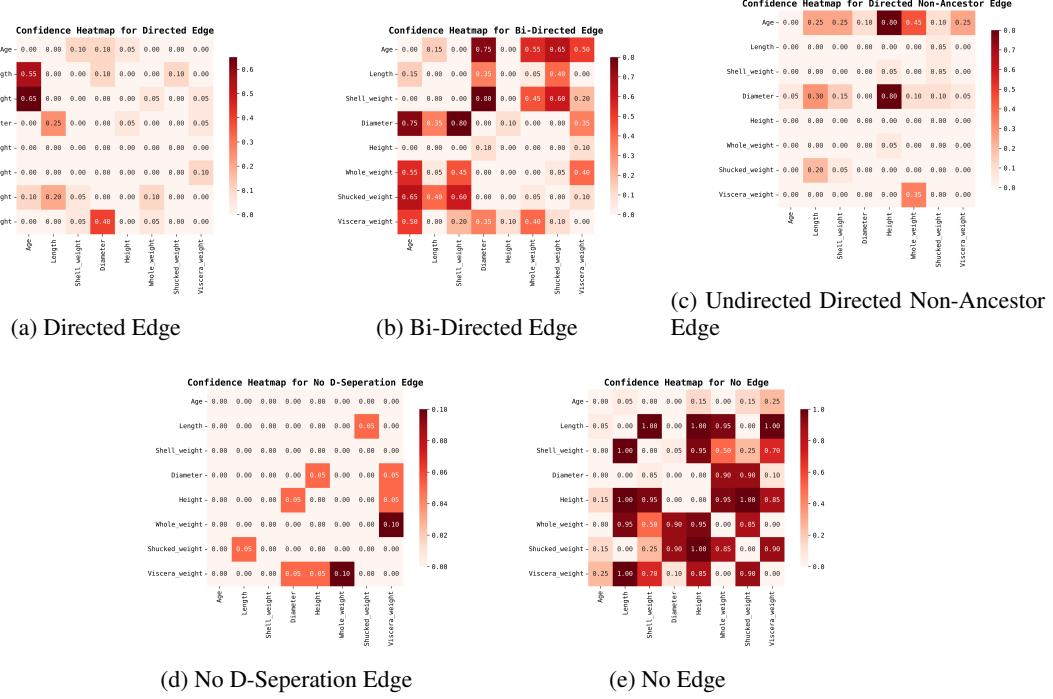


Figure 6: Confidence Heatmap of Different Edges.

The above heatmaps show the confidence probability we have on different kinds of edges, including directed edge ( $\rightarrow$ ), edge with hidden confounders ( $\leftrightarrow$ ), edge of non-ancestor ( $o \rightarrow$ ), edge of no D-Separation set ( $o - o$ ), No Edge, The heatmap of uncertain-edges, half-uncertain-edges is not shown because probabilities of all edges are 0.

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:** Age has no D-separation set with Whole weight, Age has no D-separation set with Viscera weight, Age has no D-separation set with Diameter
- **Low Confidence Edges:** Age has no D-separation set with Length, Age has no D-separation set with Height, Shell weight has no D-separation set with Height, Length has no D-separation set with Whole weight, Diameter has no D-separation set with Height, Diameter has no D-separation set with Whole weight, Diameter has no D-separation set with Shucked weight, Length has no D-separation set with Viscera weight, Age has no D-separation set with Shucked weight, Age has no D-separation set with Shell weight, Length has no D-separation set with Height, Length has no D-separation set with Shucked weight

### 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.

LLM doesn't forbid any edges.

The following are directions added by the LLM:

- **Age → Viscera weight:** As an organism ages, metabolic processes and growth patterns change, which may lead to variations in the physiological characteristics such as Viscera weight, reflecting the direct influence of aging on body composition;
- **Shell weight → Height:** Shell weight in bivalves like abalones is influenced by their growth and overall size, which includes height; hence, an increase in height through growth interventions would lead to an increase in shell weight as the organism develops.
- **Shell weight → Viscera weight:** Shell weight and viscera weight are related as both are components of the organism; as shell growth occurs, it is logical that viscera weight would also increase due to the growth of internal tissues and organs.
- **Length → Shell weight:** The length of an organism, such as a mollusk or other shelled animal, plays a significant role in determining its shell weight, as larger organisms typically have thicker and heavier shells to support their size;
- **height → shucked weight:** The height of an organism, such as an abalone, can be indicative of its maturity and overall growth, which could influence the amount of shucked weight as it is likely to be proportionate to its size;

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

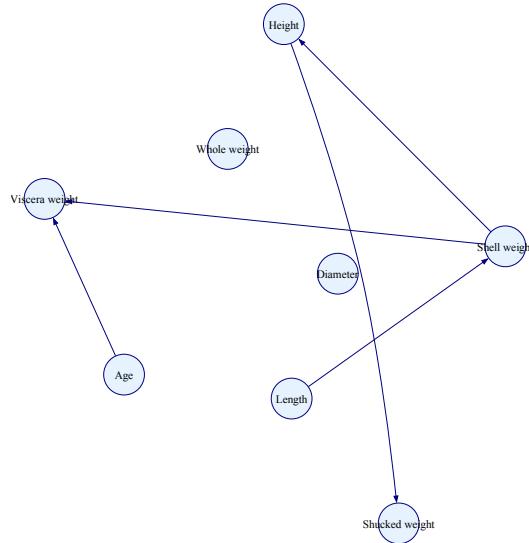


Figure 7: Revised Graph by LLM.

### 5.3 Graph Reliability Analysis

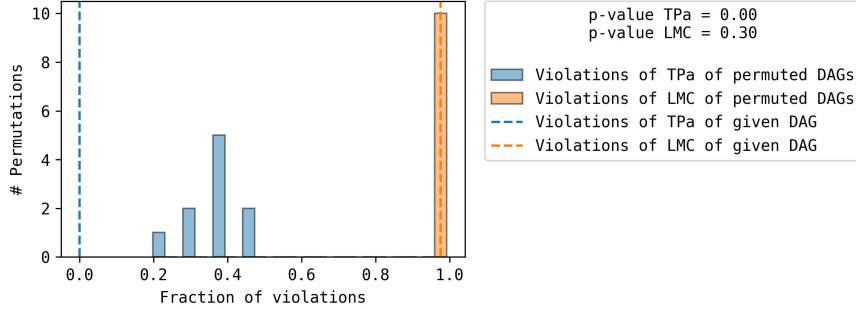


Figure 8: Refutation Graph.

The results of the graph refutation test indicate that the proposed directed acyclic graph (DAG) is unlikely to be a reliable representation of the underlying causal structure, as it exhibits significant discrepancies with the data. With a p-value of 0.00, none of the tested permutations fit within the Markov equivalence class of the DAG, which strongly suggests that the causal relationships posited by the graph do not hold. Furthermore, the DAG violates 41 out of 42 local Markov conditions (LMCs), while only surpassing 70.0% of permuted DAGs in terms of compliance with these conditions (p-value: 0.30), indicating poor alignment with expected causal constraints. Given that the results firmly reject the DAG at a significance level of 0.05, we conclude that the graph's implied causal structure is not supported by the data, warranting its rejection for any causal inference or decision-making processes based on this model.

### 5.4 Result Graph Comparision

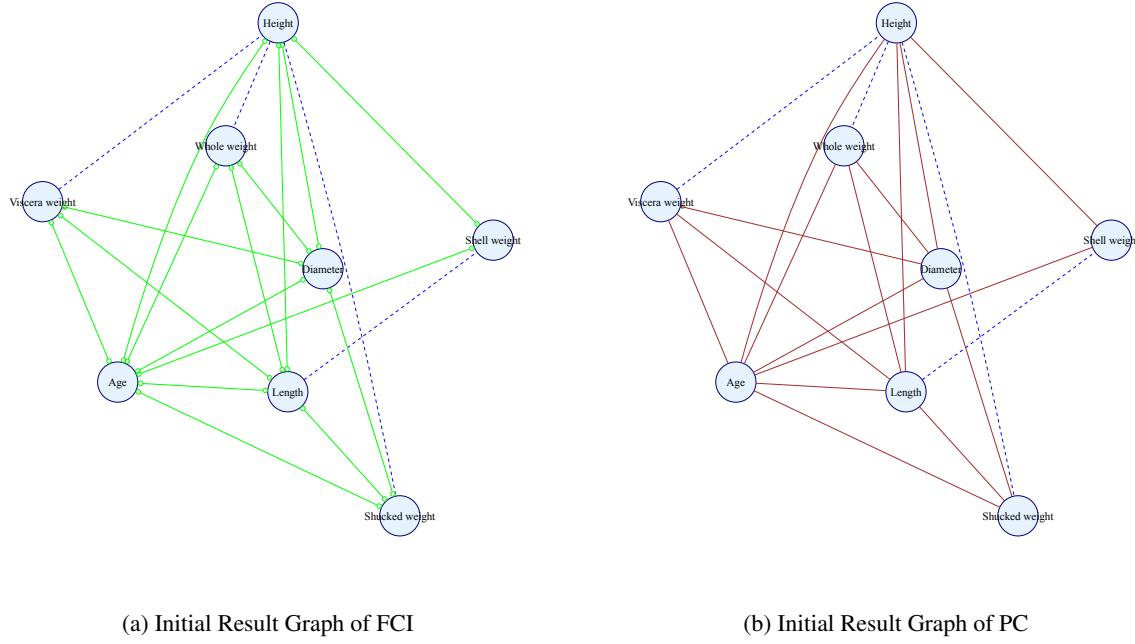


Figure 9: Result Graph Comparision of Different Algorithms.

The causal discovery results from the FCI and PC algorithms highlight several edges within their inferred causal graphs, with both showing a notable overlap in relationships among variables. The FCI results indicate that Age has no D-separation with variables such as Length, Viscera\_weight, Height, and others, suggesting that these variables

are directly related to Age without any blocking set in-between. In contrast, the PC algorithm describes the same relationships as undirected relationships, indicating that while the causal direction is not specified, the interactions among these variables are still significant.

The common edges between the two algorithms include relationships involving Age with Length, Viscera\_weight, Height, Diameter, Shell\_weight, and Whole\_weight, among others. This consistency across both results implies a strong association between these variables, making them candidates for further analysis in understanding their interdependencies.

However, when it comes to reliability, the edges presented by the FCI algorithm may be regarded as more definitive due to the explicit identification of D-separation, which suggests a clearer predictive structure within the causal relationships. The undirected relationships from the PC algorithm, while consistent, may introduce ambiguity regarding the directionality of causation. Thus, edges indicating direct relationships without D-separation in FCI could potentially be more reliable for establishing causal inferences in further studies, as they indicate a stronger connection and practical implications for interventions or predictions.

## 5.5 Conclusion

In this study, we conducted a thorough causal discovery analysis on the Abalone dataset, which includes critical biological and morphological variables such as Age, Length, Shell Weight, Diameter, Height, Whole Weight, Shucked Weight, and Viscera Weight. We employed advanced methodologies such as the Fast Causal Inference (FCI) algorithm, along with a hybrid approach (IAMBnPC), and utilized a large language model (LLM) for algorithm selection and hyperparameter optimization, followed by rigorous graph tuning with Bootstrap validation. Our findings revealed that Age is a central variable significantly impacting various metrics related to the abalones' physical growth, demonstrating direct and indirect causal relationships with Length, Shell Weight, Diameter, and Whole Weight while indicating high interdependence among these factors.

The contribution of this research lies in accurately mapping out the causal structure among these attributes and highlighting the nuances of their interrelationships. Notably, our analysis revealed that Height serves as a primary driver for Whole Weight, supplemented by Viscera Weight and Age, underscoring the complexity in their interactions. This work not only enhances understanding of the biological dynamics within abalone populations but also provides actionable insights that can inform effective management and agricultural strategies, paving the way for future research to explore further causal dynamics and potential interventions in abalone growth methodologies.

## 6 Causal Inference Results

### 6.1 Proposal Overview

In this report, we aim to delve into the intricate relationships between various features and Whole Weight, with the primary objective of identifying the key factors that influence this critical measure. By employing causal inference methodologies, we seek to unravel the complex web of dependencies and interactions among the variables contributing to Whole Weight. This analysis not only allows us to rank the most influential variables but also enhances our understanding of the underlying mechanisms at play. Such knowledge is pivotal for future analyses, as it directs subsequent research efforts and informs effective decision-making based on the recognized contributions of each factor.

With a focus on elucidating the drivers of Whole Weight, our investigation is designed to provide actionable insights that can be utilized in various applications, ranging from agricultural practices to product development and quality control. By discerning the influential factors, stakeholders can make informed choices that optimize outcomes, improve product quality, or enhance operational efficiency. This proposal emphasizes the importance of understanding causal relationships rather than mere correlations, as our goal is to identify true drivers that may yield meaningful interventions and enhancements in processes related to Whole Weight.

### 6.2 Feature Importance Analysis

#### 6.2.1 Estimation Method & Justification

**Model Type:** Random Forest

**Reasons:** We use the SHAP (SHapley Additive exPlanations) method to explain the feature importance of our Random Forest model because it provides consistent and interpretable values that reflect each feature's contribution to the model's predictions. By attributing the output to individual features based on their marginal contributions, SHAP

allows us to gain insights into the most influential factors driving Whole Weight, facilitating better decision-making and understanding of the underlying data relationships.

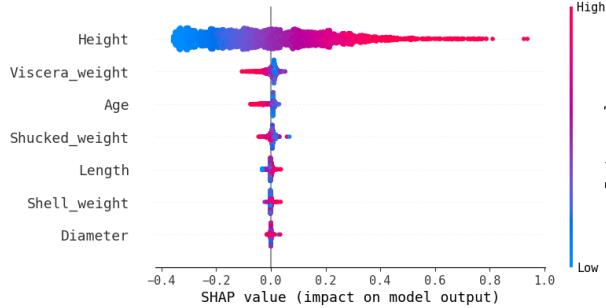


Figure 10: Beamplot of SHAP Value.

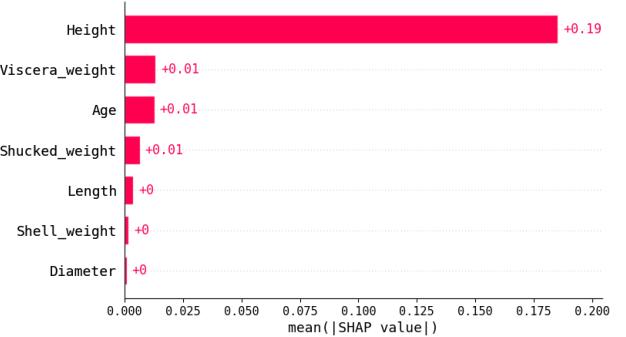


Figure 11: Barplot of Average SHAP Value.

In the context of analyzing the factors influencing Whole Weight, we have employed SHAP (Shapley Additive Explanations) values to quantify the contribution of each feature. Below are the key findings:

- **Height** (Mean SHAP Value: **0.185364**): The most influential factor affecting Whole Weight, indicating a strong positive correlation. A significant increase in Height positively impacts the Whole Weight.
- **Viscera Weight** (Mean SHAP Value: **0.013313**): This feature also shows a notable influence on Whole Weight, suggesting it plays an important role in determining overall weight, albeit less significant than Height.
- **Age** (Mean SHAP Value: **0.012859**) and **Shucked Weight** (Mean SHAP Value: **0.006579**): Both factors contribute positively to Whole Weight, indicating relevance in the analysis albeit with lesser influence compared to the top features.
- **Length** (Mean SHAP Value: **0.003746**) and **Diameter** (Mean SHAP Value: **0.000980**): These features have minimal impact on Whole Weight, suggesting they are less critical in driving this variable.
- **Shell Weight** (Mean SHAP Value: **0.001709**): Among the analyzed features, Shell Weight demonstrates the least influence on Whole Weight.

The analysis highlights that **Height** is the primary driver of Whole Weight, with **Viscera Weight** and **Age** also having positive contributions. Other features like Length, Diameter, Shell Weight, and Shucked Weight contribute less significantly. This insight can inform further research and guide potential interventions for weight optimization.

### 6.3 Summary & Next Steps

#### 6.3.1 Discussion

In this analysis, we employed **SHAP (Shapley Additive Explanations)** values to assess the factors influencing Whole Weight. SHAP values are particularly effective in quantifying the contribution of each feature, providing clarity on the importance of various variables in determining Whole Weight. This method allows us to decompose the prediction into the individual impact of each predictor, thereby enhancing our understanding of how each factor contributes to the overall outcome.

Our findings indicate that Age positively contributes to Whole Weight, with a Mean SHAP Value of **0.012859**. This suggests that while Age indeed plays a role in increasing Whole Weight, its influence is less significant in comparison to other factors such as Height and Viscera Weight.

- Overall, this highlights the nuanced interplay of different variables in shaping Whole Weight, under-scoring the importance of a comprehensive analysis in causal inference tasks.

#### 6.3.2 Next Steps Suggestions

To build on the findings from the Feature Importance analysis, a potential next step involves conducting a more focused exploration of the interactions among features that significantly impact Whole Weight. **Incorporating interaction terms in the modeling could refine our understanding of how Height, Viscera Weight, and Age work together**

**to influence Whole Weight.** Additionally, using advanced techniques such as Interaction SHAP values or Partial Dependence Plots could provide deeper insights into the non-linear relationships that may exist among these key variables. Furthermore, it may be beneficial to standardize the data by scaling features to mitigate the influence of outliers, which can skew SHAP values and result interpretations.

In terms of future research directions,

- **expanding this analysis to include external datasets for validation would enhance the robustness and generalizability of the findings.**
- Utilizing other populations or environments can help discern if the identified causal relationships hold in diverse scenarios.
- Additionally, implementing longitudinal studies would enable us to observe how changes over time in height or Viscera Weight affect Whole Weight, ultimately allowing for a more dynamic understanding of these relationships.
- **Finally, machine learning techniques combined with causal inference methods could be explored to identify potential confounders and improve the precision of our causal claims.**