# Causal Discovery Report on Abalone

Technical Report

Causal Copilot

November 4, 2024

## Abstract

This study analyzes the Abalone dataset, which encapsulates various biological and physical characteristics of abalones, to uncover causal relationships among these variables. We employed a comprehensive causal discovery methodology utilizing three algorithms—PC, GES, and NOTEARS—guided by the insights from a large language model for algorithm selection and hyperparameter tuning. The results revealed complex interdependencies, indicating that Age significantly influences Diameter and Viscera weight, while Length affects both Shell and Shucked weights, emphasizing the interconnectedness of physical attributes in organism growth. Our contributions lie in enhancing causal discovery techniques by utilizing advanced machine learning tools, thereby providing a robust framework for accurately interpreting causal structures in biological data, despite some edges demonstrating low statistical confidence. Further, our findings prompt deeper exploration into causal relationships in abalone growth patterns, integrating expert biological knowledge with statistical analysis for improved reliability.

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

## 1 Introduction

The Abalone dataset provides a rich source of biological and physical characteristics of abalones, marine mollusks known for their unique shell structure. Each variable in the dataset—such as Age, Length, Shell weight, Diameter, Height, Whole weight, Shucked weight, and Viscera weight—offers insights into the life and growth of these organisms. The Age of an abalone, determined by intricate ring structures on their shells, serves as a fundamental metric influencing various physical attributes, as older abalones typically exhibit greater Length, Diameter, and Height. Moreover, the relationships among different weight measurements reveal critical aspects of their maturity and health. Additionally, understanding the environmental factors impacting growth rates and reproductive cycles, along with potential confounding variables, is crucial for accurately deciphering causal relationships within the data. By leveraging this domain knowledge, the causal discovery process can be enhanced, providing a clearer picture of how these variables interact within the biological framework of abalones.

## 2 Background Knowledge

### 2.1 Detailed Explanation about the Variables

The variables in the Abalone dataset are meaningful and relate to the biological characteristics of abalones, which are marine mollusks. Here's a detailed overview:

- **Age**: This is typically expressed as the number of rings counted on the shell, which is a standard method for estimating the age of abalones. Each ring corresponds to one year of growth.
- **Length**: This measures the overall length of the abalone's body (or shell) from the apex to the bottom, typically expressed in millimeters.
- **Shell weight**: This is the weight of the abalone's shell, measured in grams. It reflects the maturity and health of the abalone.

- **Diameter**: This represents the measurement across the widest part of the shell, also typically in millimeters.
- **Height**: This is the vertical measurement from the bottom of the shell to the top, measured in millimeters.
- **Whole weight**: This is the total weight of the abalone, including the shell and soft tissue, measured in grams. It gives an indication of the abalone's size and health.
- **Shucked weight**: This refers to the weight of the abalone without the shell, measuring just the soft tissue, also in grams.
- **Viscera weight**: This is the weight of the internal organs of the abalone, measured in grams, which can provide insights into its health and reproductive status.

### 2.2 Possible Causal Relations among these Variables

- **Age** → **Length**: Generally, as abalones age, they tend to grow longer. Therefore, Age may causally affect Length.
- **Age** → **Diameter**: Similar to Length, as abalones increase in Age, they usually also increase in Diameter.
- **Age** → **Height**: Age may influence Height, with older abalones usually being taller as they grow.
- **Length** → **Shell weight**: Longer abalones are typically heavier in terms of their Shell weight, indicating a causal relation based on physical growth.
- **Diameter** → **Shell weight**: Wider abalones likely exhibit a higher Shell weight, reflecting the correlation with their size.
- **Whole weight** → **Shucked weight**: As the total body size (Whole weight) increases, the portion that is only flesh (Shucked weight) also increases.
- **Whole weight** → **Viscera weight**: An increase in Whole weight may indicate larger internal organs, suggesting a causal relationship here.
- **Shucked weight** → **Viscera weight**: As the edible part of the abalone increases in weight, it may correlate positively with the weight of the Viscera.

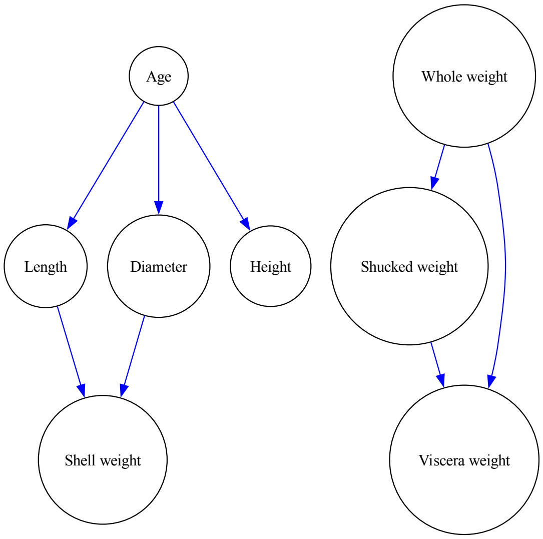


Figure 1: Possible Causal Relation Graph

# 3 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview

| Age | Length | Shell weight | Diameter | Height | Whole weight | Shucked weight | Viscera weight |
|---|---|---|---|---|---|---|---|
| 15.0 | 0.455 | 0.365 | 0.095 | 0.5140 | 0.2245 | 0.1010 | 0.150 |
| 7.0 | 0.350 | 0.265 | 0.090 | 0.2255 | 0.0995 | 0.0485 | 0.070 |
| 9.0 | 0.530 | 0.420 | 0.135 | 0.6770 | 0.2565 | 0.1415 | 0.210 |
| 10.0 | 0.440 | 0.365 | 0.125 | 0.5160 | 0.2155 | 0.1140 | 0.155 |
| 7.0 | 0.330 | 0.255 | 0.080 | 0.2050 | 0.0895 | 0.0395 | 0.055 |

## 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 |
|---|---|---|---|---|---|---|
| (4177, 8) | 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.

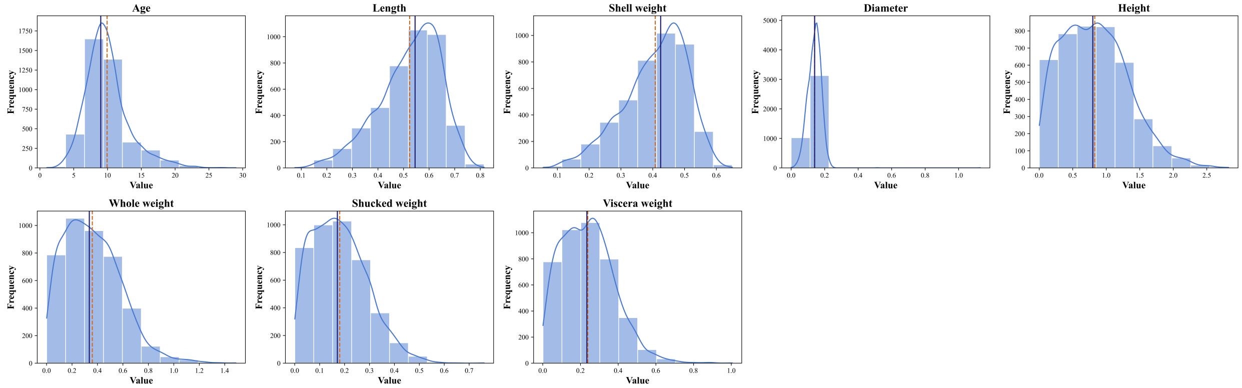


Figure 2: Distribution Plots of Variables

- Slight left skew distributed variables: Length, Shell Weight, Diameter, Whole Weight
- Slight right skew distributed variables: Age, Height, Shucked Weight, Viscera Weight
- 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: Shell weight and Length, Height and Whole weight, Shucked weight and Height, Viscera weight and Height
- Moderate Correlated Variables: Length and Age, Shell weight and Age, Diameter and Age, Height and Age, Whole weight and Diameter, Shucked weight and Diameter
- Weak Correlated Variables: None

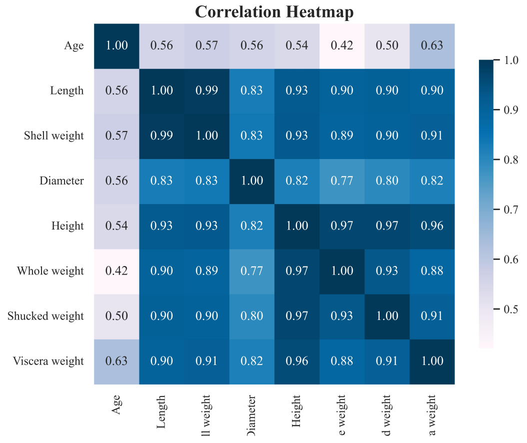


Figure 3: Correlation Heatmap of Variables

## 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**: This algorithm is suitable due to the large sample size (4177) and the absence of missing values, allowing for reliable statistical independence tests. PC operates efficiently with large-scale data like this, making it a good option.
- **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 because it can handle non-Gaussian data and provides a more direct approach to finding the best causal structure among the observed variables. The dataset's characteristics indicate non-linearity, which GES can efficiently manage.
- **NOTEARS**:
  - **Description**: NOTEARS transforms the problem of learning Directed Acyclic Graphs (DAGs) into a continuous optimization problem, allowing for efficient scaling to large datasets.
  - **Justification**: Due to its adaptability for potentially high-dimensional data and efficiency, NOTEARS is a viable option. While it assumes linearity, its optimization framework may still yield valuable insights in exploring the causal structure of the dataset.

### 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**: This value is appropriate as the sample size is 4177, which falls within the range of 500-10,000. Therefore, a significance level of 0.05 balances the risk of Type I errors while preserving robustness in conditional independence tests.
- **indep_test**:
  - **Value**: fisherz
  - **Explanation**: Given that the data is continuous, the Fisher's Z test is suitable. Although the dataset does not exhibit predominant linearity or Gaussian errors, the absence of missing values allows for the application of this method to perform independence tests effectively.
- **depth**:
  - **Value**: -1
  - **Explanation**: With 8 features in the dataset, setting the depth to -1 allows for unlimited search depth, which is beneficial for capturing possible causal structures accurately. Speed optimization is less of a priority in this context given the small number of nodes.

### 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.

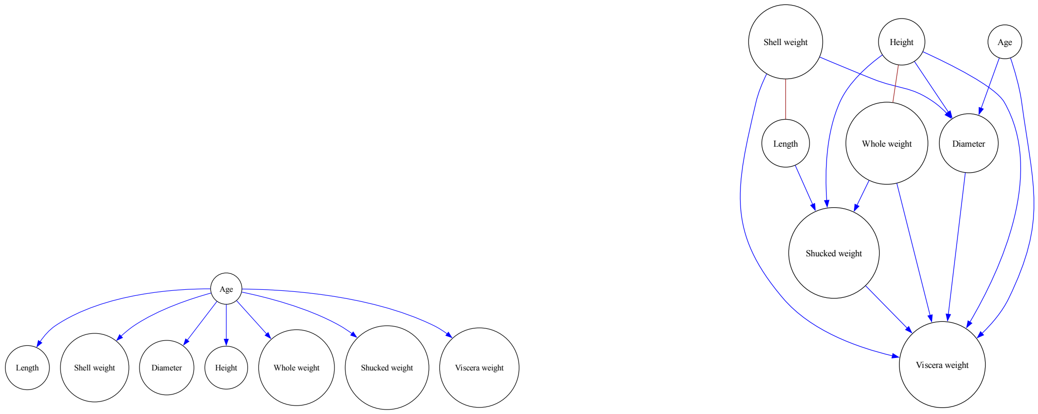
Firstly, we use the Bootstrap technique to get how much confidence we have on each edge in the initial graph. If the confidence probability of a certain edge is greater than 95% and it is not in the initial graph, we force it. Otherwise, if the confidence probability is smaller than 5% and it exists 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 Bootstrap and LLM to refine the causal graph, we can achieve improvements in graph 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 reveals a complex web of causal relationships among the variables, highlighting how Age influences multiple factors within a biological context. Specifically, Age is a determinant of both Diameter and Viscera weight, suggesting that as organisms mature, their body size and internal organ weight increase. Length plays a critical role as it not only causes Shell weight but also Shucked weight, indicating that as the organism grows longer, both its shell and the edible meat yield increase. Shell weight further impacts Diameter and Viscera weight, emphasizing the interconnectedness of physical attributes in relation to body composition. Additionally, Height is positioned as a causal precursor to Diameter, Whole weight, Shucked weight, and Viscera weight, underlining the significance of overall height in the developmental aspects of the organism. This cascading effect of one variable influencing several others illustrates how these physical characteristics are intricately linked in shaping the organism's overall growth and development.

### 5.2 Revised Graph

By using the method mentioned in 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:

- **Age → Length**: As abalones age, they generally grow longer, indicating a direct causal relationship where an increase in Age leads to an increase in Length.
- **Age → Shell weight**: Older abalones typically have a heavier shell due to increased size and maturation, showing that Age causally affects Shell weight.
- **Age → Height**: With aging, abalones usually become taller; thus, Age can be seen as a factor that causally influences the Height of abalones.
- **Age → Whole weight**: The overall size and weight of an abalone increase with Age due to growth, establishing a causal link from Age to Whole weight.
- **Age → Shucked weight**: As abalones age and grow, their edible flesh increases in weight, suggesting Age causally impacts Shucked weight.
- **Length → Diameter**: There is a typical growth relationship where an increase in Length correlates with an increase in Diameter, indicating that Length causally affects Diameter.
- **Length → Height**: Longer abalones usually also display increased Height, establishing a causal relationship where Length affects Height.
- **Length → Whole weight**: Increased Length typically contributes to a higher overall body mass, suggesting that Length causally influences Whole weight.
- **Length → Viscera weight**: As the Length of an abalone increases, it tends to have larger internal organs, indicating a causal effect from Length to Viscera weight.
- **Shell weight → Shucked weight**: The weight of the shell impacts the amount of flesh available, meaning that an increase in Shell weight could causally lead to an increase in Shucked weight.

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

- **Length → Shell weight**: Longer abalones are usually heavier, meaning that Length directly affects the Shell weight as size increases.

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

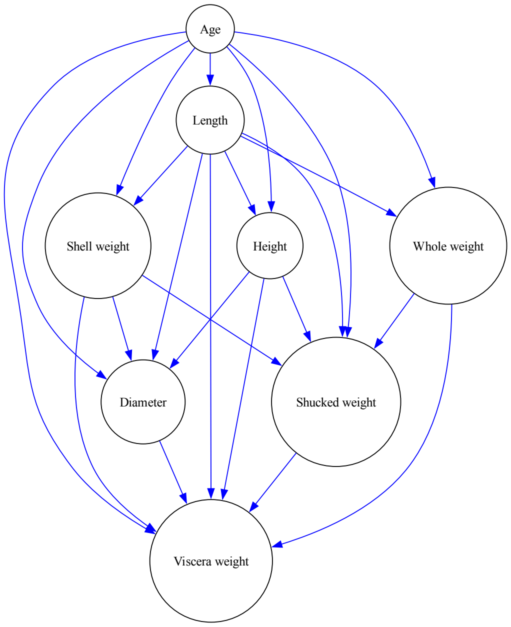
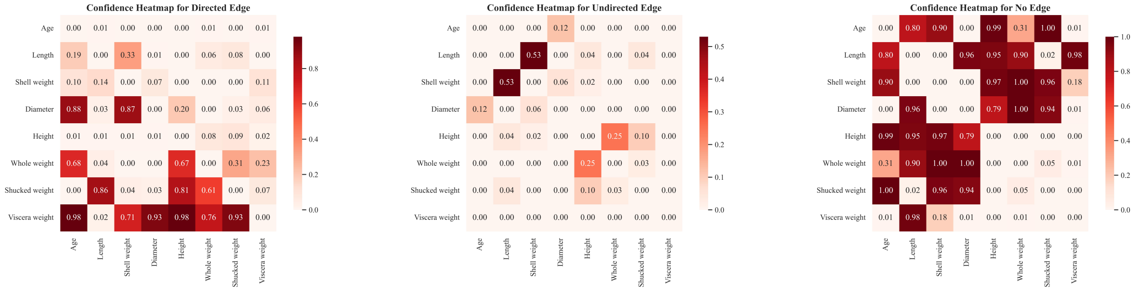


Figure 5: Revised Graph

### 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 statistics perspective, we have high confidence to believe that the edges **Whole weight → Height (0.67)** and **Whole weight → Shucked weight (0.31)** exist, indicating a significant causal relationship as larger abalones typically show greater Height and Shucked weight. Edges like **Age → Diameter (0.0)** and **Age → Viscera weight (0.01)** demonstrate a lack of confidence, suggesting that these relationships are unlikely to be true in the dataset.

However, based on expert knowledge, we know that certain edges are reasonable to presume exist, such as **Age → Length**, **Age → Diameter**, and **Length → Shell weight**, which all align with biological understanding of abalone growth patterns. Conversely, edges like **Height → Diameter (0.01)** and **Shell weight → Diameter (0.07)** may not hold as strongly in practice, since Diameter is more directly related to other physical dimensions rather than Height and Shell weight alone.

Therefore, while some relationships in the causal graph are supported statistically, the low bootstrap probabilities of several edges and the complexity of the biological interactions suggest that the result of this causal graph is not fully reliable without further validation.