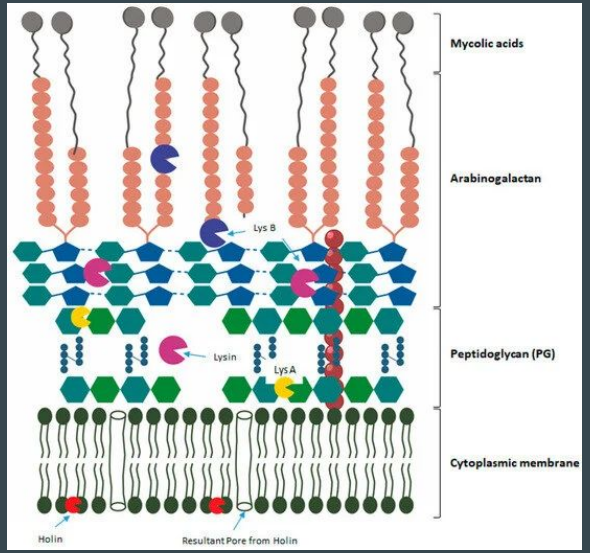
Identifying Top Lysin B Enzymes for Phage Therapy

Exploring Biochemical Properties and Sequence Conservation to Optimize Purification and Efficacy

What is Lysin B and its role in bacteriophage therapy?

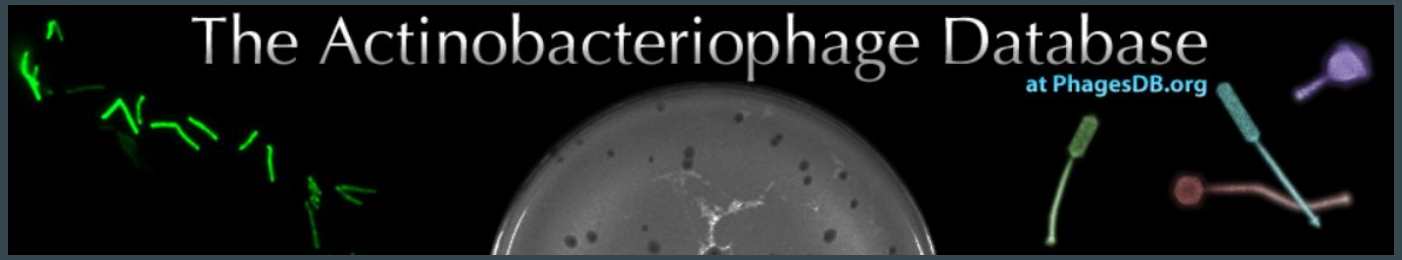
- Lysin B
  - Produced by bacteriophages and degrades the mycolic acid layer of bacterial cell walls, particularly in Mycobacteria
- Therapeutic Applications
  - Effective lysins must be carefully selected for their ability to target bacteria while being easy to purify for scalable production.
  - These lysins can enhance other treatments, such as phages or antibiotics, through a combination of computational predictions and experimental validation.



What are the Project Goals and Objectives?

- Goals:
  - Identify top-performing Lysin B enzymes by analyzing biochemical properties and sequence conservation across various bacteriophages.
  - Prioritize enzymes based on ease of purification and therapeutic potential against antibiotic-resistant Actinobacteria, particularly Mycobacteria.
- Objectives
  - Identify easily purifiable Lysin B enzymes, focusing on physicochemical properties.
  - Discover sequence conservation among Lysin B enzymes and phages.
  - Develop machine learning models to predict optimal Lysin B variants with desired properties and conservation
  - Guide the selection of effective Lysin B enzymes and corresponding phages for therapeutic use against Mycobacteria infections.

What Data are We Using?



- Data Source:
  - The primary data was fetched from the PhagesDB API: https://phagesdb.org/api/genes/
  - PhagesDB is a online database dedicated to bacteriophages, including their genomic sequences, structural features, and biological characteristics.
- Data Preprocessing
  - The Notes column was parsed in a Pandas DataFrame to identify entries related to Lysin B.
  - Rows containing "Lysin B" in the Notes field were filtered and saved into a new DataFrame.
  - Calculate a Treatment Score using Biochemical and Sequence Conservation Properties

What Additional Features are Important for our Analysis?

- Molecular Weight: Total mass; important for size and behavior.
- Isoelectric Point (pI): pH with no net charge; affects solubility and purification.
- Hydrophobicity: Water repulsion; impacts interaction with bacterial membranes.
- Aliphatic Index: Indicates thermostability; higher values suggest greater stability.
- Aromaticity: Proportion of aromatic amino acids; influences structural stability.
- Instability Index: Predicts stability; lower values indicate higher stability.
- Protein Length: Total amino acids; affects function and interactions.
- Multiple Sequence Alignment (MSA): Identifies conserved regions of functional importance.
- Consensus Sequence: Represents the most common amino acids at each position.
- Hamming Distance: Measures similarity to the consensus sequence.

How do we Calculate Treatment Score?

- Score Categories:
  - Stability: Instability Index, Isoelectric Point, Aliphatic Index, Aromaticity.
  - Hydrophobicity: Gravy Score and Hydrophobicity Score.
  - Size: Molecular Weight and Protein Length.
  - Similarity: Hamming Distance from consensus.
- Final Treatment Score:
  - Derived by averaging the four group scores (Stability, Hydrophobicity, Size, Similarity) to rank Lysin B proteins based on stability, solubility, size, and conservation suitability for therapeutic development.

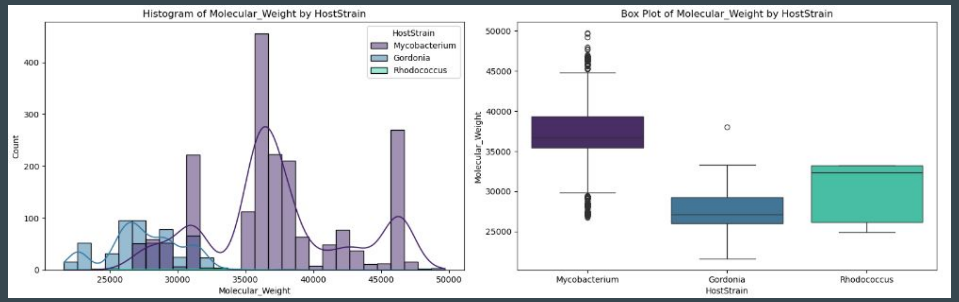
Exploratory Analysis of Lysin B Protein Features and Treatment Score Correlation

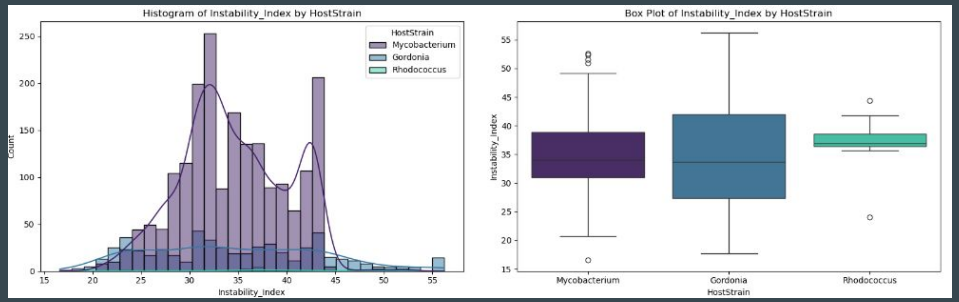
- Explore distribution and variability of Lysin B protein features across different Host Strains.
  - Created histogram and box plots for biochemical features across different host strains
- Examine how scaled and encoded features correlate with the Treatment Score of each Lysin B protein.
  - Created scatter plots for each feature to visualize relationships with the Treatment Score.
  - Assessed whether relationships were linear or non-linear.

Visualization of Biochemical Properties by Host Strain

- Histograms:
  - Displayed distribution of values for each feature.
  - Kernel Density Estimate (KDE) overlaid to highlight probability density, differentiated by Host Strain using color coding.
- Box Plots:
  - Illustrated variability of each feature across different host strains.
  - Enabled comparison of spread and potential outliers in Lysin B protein properties.

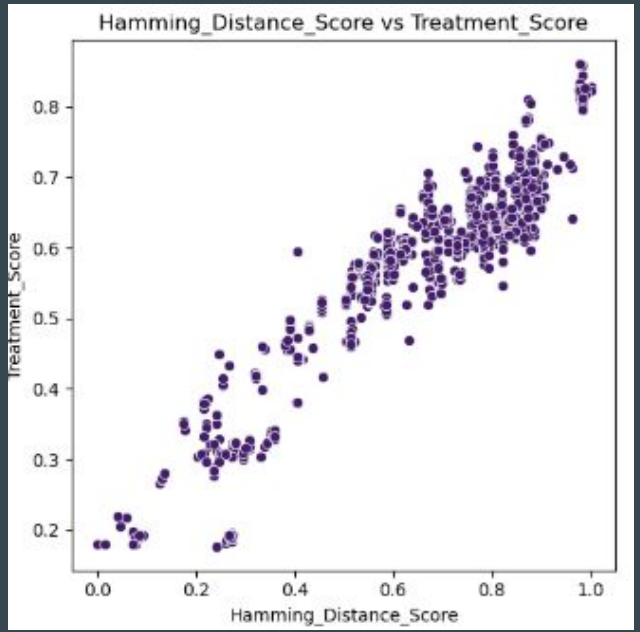
Some features exhibited variability between host strains, while others remained more uniform.

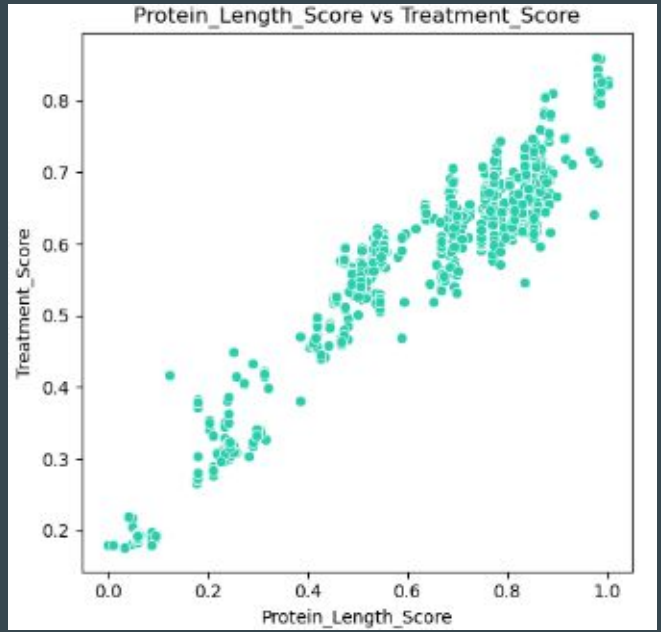




Strong Positive Correlation of Features with Treatment Score

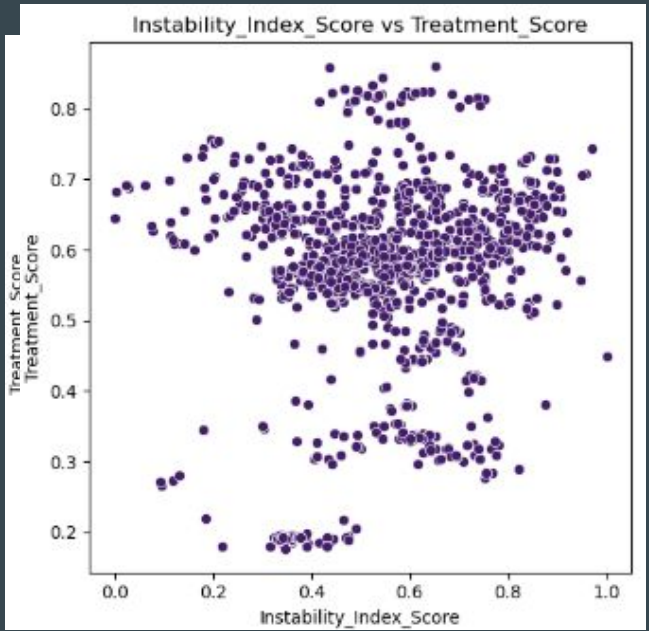
Scatter plots illustrated how features influence the Treatment Score.

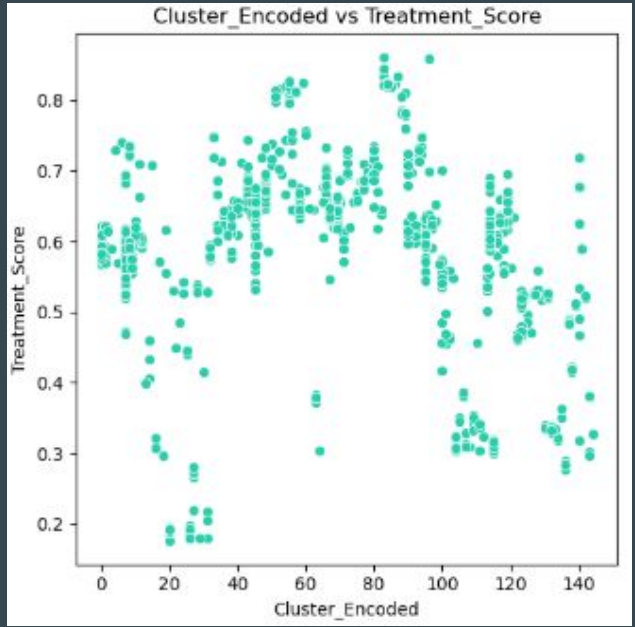




Weak/Complex Correlation of Features with Treatment Score

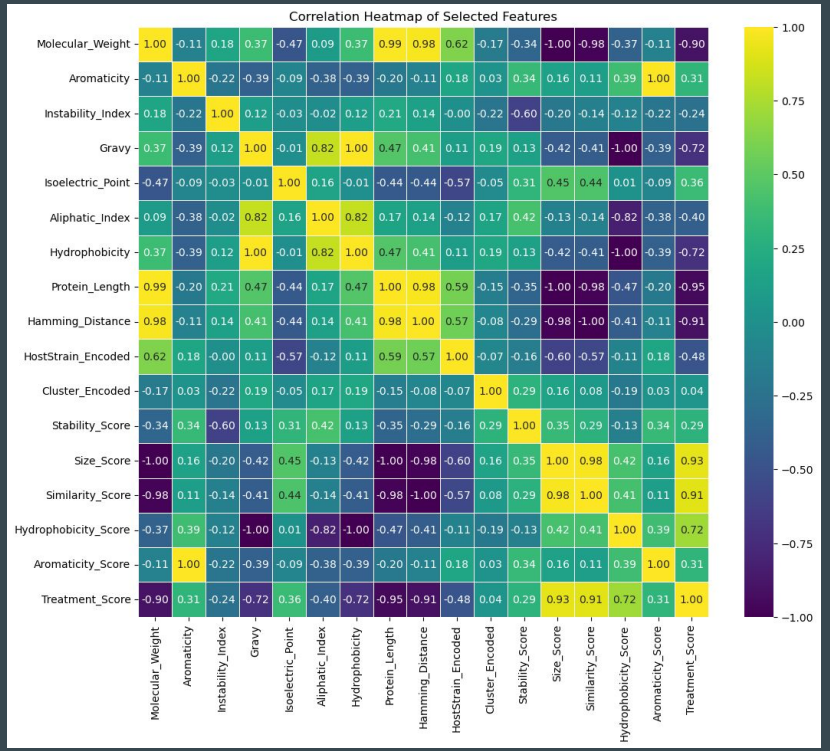
Scatter plots illustrated how features influence the Treatment Score.





Correlation Heatmap Features vs. Treatment Score

- Quantify relationships between various features and the Treatment Score.
- Pearson Correlation Coefficients were computed for all features using a Heatmap Correlation Matrix
- Strong Correlation
  - Hydrophobicity Score
  - Protein Length
- Weak Correlation
  - Cluster Encoded
  - Instability Index



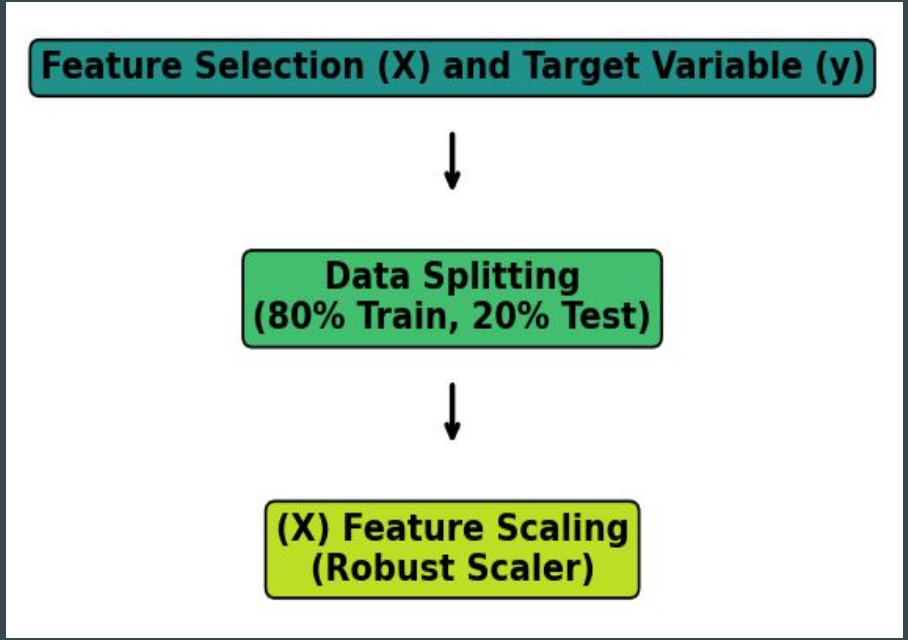
Model Preprocessing

Feature Selection (X)

- Isoelectric Point
- Aromaticity
- Instability Index
- Aliphatic Index
- Gravy
- Hydrophobicity
- Molecular Weight
- Protein Length
- Hamming Distance
- HostStrain Encoded

Target Variable (y)

- Treatment Score



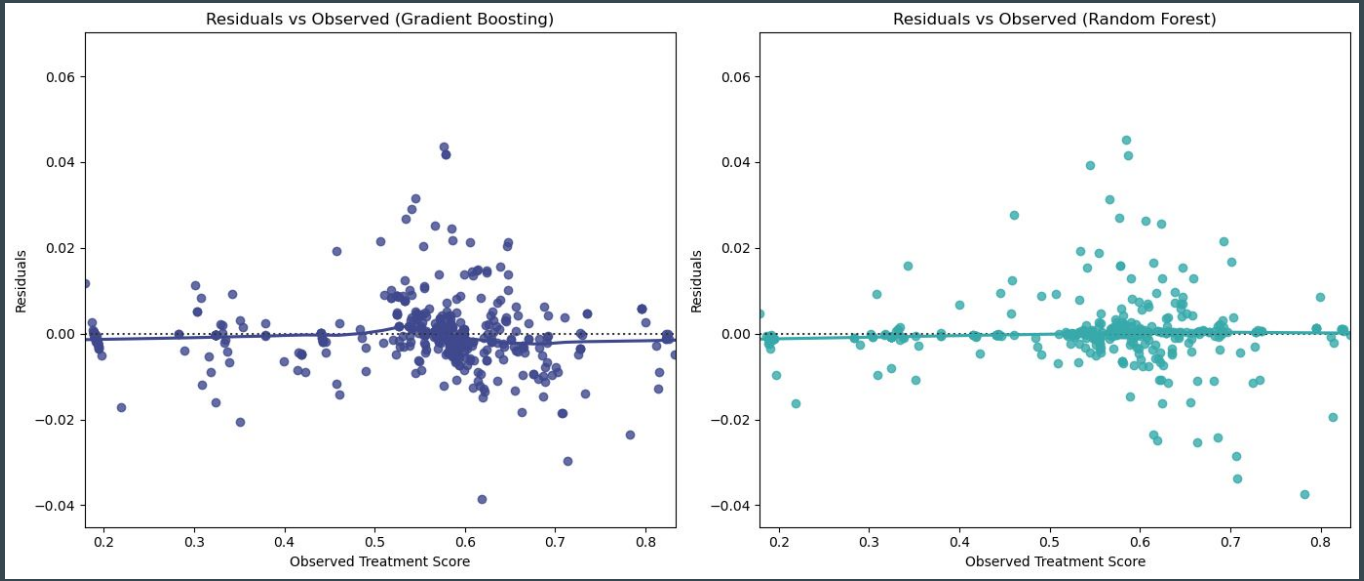
Model Selection, Training, and Performance Metrics

- Evaluated multiple machine learning models for Treatment Score predictions
  - Closer to 0: better performance:
    - MSE, RMSE, MAE, MAPE
  - Closer to 1: better performance:
    - R^2, Explained Variance

Model	Train MSE	Test MSE	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2	Train Explained Variance	Test Explained Variance	Train MAPE	Test MAPE
Linear Regression	0.000078	0.000071	0.006125	0.006018	0.008846	0.008440	0.996767	0.996808	0.996767	0.996810	0.011673	0.011505
Ridge Regression	0.000078	0.000071	0.006136	0.006040	0.008847	0.008455	0.996766	0.996797	0.996766	0.996799	0.011722	0.011572
Random Forest	0.000011	0.000058	0.001140	0.002994	0.003256	0.007599	0.999562	0.997413	0.999562	0.997413	0.002105	0.005359
Gradient Boosting	0.000033	0.000058	0.003608	0.004509	0.005731	0.007642	0.998643	0.997383	0.998643	0.997383	0.006929	0.008589
Support Vector Regression	0.002059	0.001971	0.035293	0.034341	0.045373	0.044394	0.914928	0.911696	0.915036	0.911725	0.098197	0.095607

Residual Analysis

- Residuals for both models are close to 0, confirming high accuracy between observed and predicted values



GridSearchCV Model Tuning and Evaluation

Random Forest:

- Best GridSearch Parameters
  - n_estimators: 300, max_depth: 20, min_samples_split: 2, min_samples_leaf: 1, max_features: 'sqrt'
- Cross Validation Scores
  - RMSE: 0.00730 ± 0.00121

Gradient Boosting:

- Best GridSearch Parameters
  - n_estimators: 300, learning_rate: 0.05, max_depth: 5, min_samples_split: 2, min_samples_leaf: 4, subsample: 1.0, max_features: 'sqrt'
- Cross Validation Scores
  - RMSE: 0.00549 ± 0.00091

- Both models perform well, with Gradient Boosting showing slightly better metrics (lower MSE, RMSE, MAE, and higher R^2).
- Proceeding with Bayesian Optimization for further tuning of the Gradient Boosting model.

Gradient Boosting with Bayesian Optimization

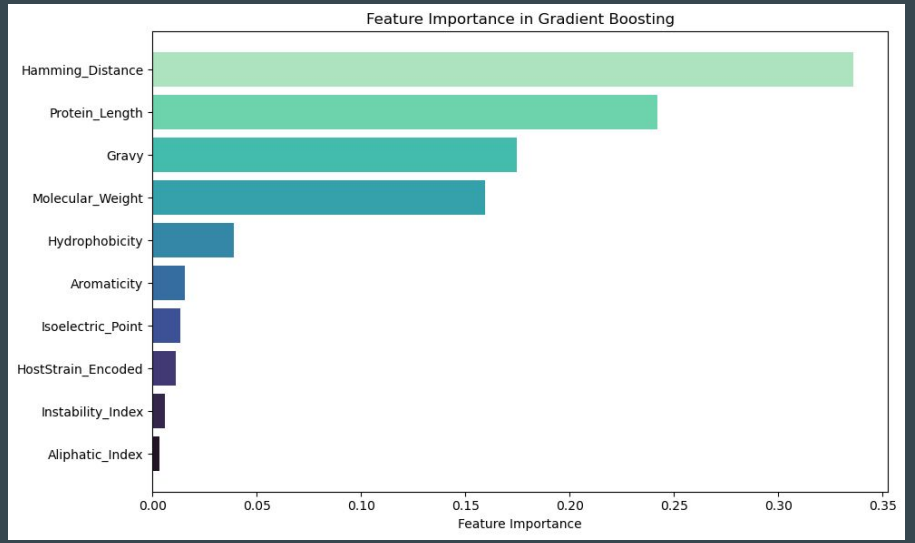
- Best Parameters:
  - n_estimators: 500
  - learning_rate: 0.075
  - max_depth: 10
  - min_samples_split: 6
  - min_samples_leaf: 8
  - subsample: 0.641
  - max_features: 'log2'

- Gradient Boosting with Bayesian Optimization further enhances model predictions, demonstrating improved accuracy across all metrics.

Model	Train MSE	Test MSE	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2	Train Explained Variance	Test Explained Variance	Train MAPE	Test MAPE
Gradient Boosting Grid Search	2.722014e-06	0.000031	0.001013	0.002619	0.001650	0.005582	0.999888	0.998604	0.999888	0.998605	0.002012	0.004977
Gradient Boosting Bayesian Optimization	1.529958e-07	0.000024	0.000199	0.001855	0.000391	0.004852	0.999994	0.998945	0.999994	0.998946	0.000406	0.003716

Feature Importance in Gradient Boosting

- Hamming Distance and Protein Length are the most influential features in predicting the Treatment Score.
- Other important features include Gravy and Molecular Weight, which also contribute significantly to model performance.



Conclusion on Model Performance and Predictive Accuracy

- Key Findings:
  - Best Model: Gradient Boosting Regressor tuned with Bayesian Optimization.
    - Lowest error metrics: MSE, RMSE, MAE.
    - Highest R^2 and Explained Variance scores.
    - Generalization: Strong performance consistency across training and test sets.
    - Mean Absolute Percentage Error (MAPE) was low, indicating close predictions to actual Treatment Scores.
- Predicted Treatment Scores:
  - Ranged from 0.177 to 0.859, indicating effectiveness and variability in lysin proteins.
    - Lower Bound (0.177): Less favorable for therapeutic efficacy and purification.
    - Upper Bound (0.859): Highly effective proteins with optimal biochemical properties.

Focus on Mycobacterium Lysin B

- Targeting Mycobacterium:
  - Specific attention to lysins for Mycobacterium strains (e.g., tuberculosis, Mycobacterium abscessus).
  - Top-performing lysins exhibited high Treatment Scores, indicating their potential for effective purification and efficacy due to their sequence conservation, allowing for treatment across multiple bacterial strains.

GeneID	HostStrain	Predicted_Treatment_Score	Treatment_Score
Kimona_CDS_5	Mycobacterium	0.740391	0.740749
Puppy_CDS_9	Mycobacterium	0.734995	0.734895
Pistachio_CDS_9	Mycobacterium	0.734995	0.734895
Idleandcovert_CDS_8	Mycobacterium	0.734995	0.734895
TNguyen7_CDS_9	Mycobacterium	0.734995	0.734895
BlueBird_CDS_9	Mycobacterium	0.734995	0.734895
Fred313_CDS_8	Mycobacterium	0.731060	0.731949
BabyRay_CDS_9	Mycobacterium	0.729006	0.728869
MA5_CDS_9	Mycobacterium	0.727330	0.726801
Phantastic_CDS_9	Mycobacterium	0.725917	0.727489

Future Project Directions

- Incorporating Additional Features:
  - Integrate biological features to enhance model accuracy:
    - Protein structural data (e.g., secondary or tertiary structure).
    - Post-translational modifications affecting stability.
    - Binding affinities between lysins and bacterial cell walls.

- Experimental Validation:
  - Validate model predictions by testing top-predicted lysins on bacterial cultures.
  - Assess ease of purification to ensure practical applications in drug development.
  - Explore combination treatments with phage therapy or antibiotics for enhanced effectiveness.

- Overall Objective:
  - Combine computational predictions with experimental validation to identify effective lysins and improve treatment options.