Identifying Top Lysin B Enzymes for Phage Therapy

Exploring Biochemical Properties and Sequence Conservation to Optimize Purification and Efficacy

Problem Statement

> This project aims to identify top-performing Lysin B enzymes by analyzing their biochemical properties and sequence conservation across various bacteriophages, with a focus on ease of purification and therapeutic potential against antibiotic-resistant Actinobacteria, particularly Mycobacteria.

Background

> Lysin B, produced by bacteriophages, degrades the mycolic acid layer of bacterial cell walls, particularly in Mycobacteria, making it a promising tool in bacteriophage therapy. While Lysin B shows great potential, its therapeutic application requires careful selection of enzyme variants that are not only effective against target bacteria but also easy to purify for scalable production.
>
> Biochemical properties such as molecular weight, isoelectric point, and hydrophobicity can impact the ease of purification and stability of Lysin B. Additionally, identifying conserved sequences in Lysin B enzymes are indicators of functional stability and efficacy across different bacterial strains, making them crucial for broad therapeutic applications.
>
> Using computational biology tools and machine learning models, this project aims to analyze the protein sequences of Lysin B enzymes across different phages. The goal is to identify enzymes with conserved sequences that can maintain efficacy while also being more straightforward to purify. These insights will help in selecting the most viable Lysin B variants and their corresponding phages for future therapeutic development.

Goal

> The primary goal of this project is to apply computational approaches to analyze Lysin B protein sequences and their associated biochemical properties in order to:
>
> - Identify Lysin B enzymes that are easiest to purify based on their physicochemical properties such as molecular weight, hydrophobicity, and isoelectric point.
> - Discover sequence conservation across different Lysin B enzymes and their corresponding phages, which may suggest higher stability, ease of production, and effectiveness in breaking down bacterial cell walls.
> - Develop machine learning models that can predict the top Lysin B variants with the desired properties, ease of purification, and sequence conservation.
> - Use these insights to guide the selection of the best Lysin B enzymes and their bacteriophages for therapeutic use, particularly against Mycobacteria infections.
>
> By achieving these goals, the project aims to streamline the identification and production of highly effective Lysin B enzymes that are both easy to purify and functionally conserved, making them more feasible for use in phage therapy.

Data Source

> The primary data used in this project was fetched from the PhagesDB API, specifically the endpoint:

URL: https://phagesdb.org/api/genes/ and saved as 'lysin_nested_data.json'

> The JSON file contains details such as GeneID, Phams, Phage Name, Protein Translations, Host Strain, Phage Cluster, and Gene Notes from various bacteriophages. The data was loaded into a pandas DataFrame named lysin_nested_data.json.

Data Preprocessing

Lysin B Extraction

> To extract relevant entries, the Notes column in the DataFrame was parsed to identify gene names or descriptions related to Lysin B. Rows containing Lysin B in the Notes field were filtered and saved into a new DataFrame, lysin_b_df, ensuring that only genes encoding Lysin B were included for further analysis.

Biochemical Properties of Lysin B Proteins

> For each Lysin B protein sequence, key biochemical properties were calculated to inform the subsequent analysis and model development. These properties include:
>
> - Molecular Weight: Represents the total mass of the Lysin B protein, which is critical for understanding its size and behavior in biological processes.
> - Isoelectric Point (pI): The pH at which the Lysin B protein carries no net charge. This property is important for understanding the protein's solubility and plays a key role in the purification process.
> - Hydrophobicity: Refers to the degree to which the Lysin B protein repels water. This property impacts the protein's interaction with bacterial membranes, which is crucial for its antibacterial efficacy.
> - Aliphatic Index: A measure related to the thermostability of the protein, calculated based on the volume occupied by aliphatic side chains. A higher aliphatic index generally indicates greater stability.
> - Aromaticity: The proportion of aromatic amino acids in the protein sequence. This can influence the structural stability and folding of the protein.
> - Instability Index: Predicts the stability of the protein. A lower instability index suggests that the protein is more stable in vitro conditions.
> - Protein Length: The total number of amino acids in the Lysin B protein, which can affect its overall function and interaction with other molecules.
>
> These properties were calculated and stored as features, forming the basis for feature engineering and the machine learning models used in later stages.

Outlier Detection

> Outliers in biochemical properties such as Molecular Weight, Aromaticity, Instability Index, and Protein Length were detected using a modified IQR (Interquartile Range) method. This helped identify extreme values that might indicate sequencing errors or uncommon variants.

Sequence Conservation Properties

> For each Lysin B protein sequence, key biochemical properties were calculated to inform the subsequent analysis and model development. These properties

include:

- Multiple Sequence Alignment (MSA) was performed to identify conserved regions, as these regions are indicative of functional importance. A consensus sequence was generated, summarizing the most common amino acids across the dataset.
- Consensus Sequence Generation: A consensus sequence was generated from the aligned sequences, representing the most common amino acids at each position.
- Hamming Distance Calculation: For each sequence, the Hamming Distance from the consensus was calculated, providing a measure of sequence similarity or divergence.

Data Integration

Biochemical and Sequence Conservation Properties were calculated and stored as individual features, forming the foundation for feature engineering and machine learning models in subsequent stages of the project.

Feature Scaling and Inversion

We applied feature scaling to normalize various biochemical properties and sequence characteristics of the Lysin B proteins. This step ensures that all features are on a comparable scale, making them suitable for subsequent analysis and model development. The MinMaxScaler from the scikit-learn library was used to scale the features between 0 and 1. Additionally, certain features were inverted where lower values indicated better stability or effectiveness.

- Instability Index: The score was inverted (1 - scaled_value) since lower instability values indicate better stability.
- Isoelectric Point: The absolute distance from pH 7 (neutral pH) was scaled, where values closer to 7 are more favorable for protein solubility.
- Aliphatic Index: Scaled without inversion, as higher values generally indicate greater stability.
- Aromaticity: Scaled directly, indicating the proportion of aromatic amino acids.
- Gravy: Inverted, as lower hydrophobicity (Gravy score) is often more favorable in aqueous environments.
- Hydrophobicity: Similarly inverted to favor lower values.
- Molecular Weight and Protein Length: Both inverted, as smaller proteins may be easier to purify and are less likely to aggregate.
- Hamming Distance: Inverted to reward sequences that are more similar to the consensus sequence, indicating conservation.

Treatment Score Calculation

> For each Lysin B sequence, the scaled features were grouped into categories representing different aspects of the protein's properties:
>
> - Stability: Includes the Instability Index, Isoelectric Point, Aliphatic Index, and Aromaticity. A higher Stability Score reflects greater protein stability, which is crucial for the enzyme's effectiveness in therapeutic applications.
> - Hydrophobicity: Combines the Gravy Score and Hydrophobicity Score, where lower scores indicate lower hydrophobicity. Lower hydrophobicity is often desirable for improved solubility in biological environments, which enhances the protein's bioavailability.
> - Size: Includes Molecular Weight and Protein Length, where smaller sizes are generally favorable for ease of purification, handling, and potentially reducing immunogenicity in therapeutic use.
> - Similarity: Measures the Hamming Distance from the consensus sequence, with a lower distance indicating a more conserved sequence. A higher Similarity Score rewards sequences that are more conserved and potentially more effective as antibacterial agents due to their evolutionary conservation.
>
> Once the group scores were calculated, the Final Treatment Score was derived by averaging the four group scores: Stability Score, Hydrophobicity Score, Size Score, and Similarity Score. This Treatment Score provides an overall ranking of each Lysin B protein, reflecting its suitability for therapeutic development based on factors like stability, solubility, size, and sequence conservation.

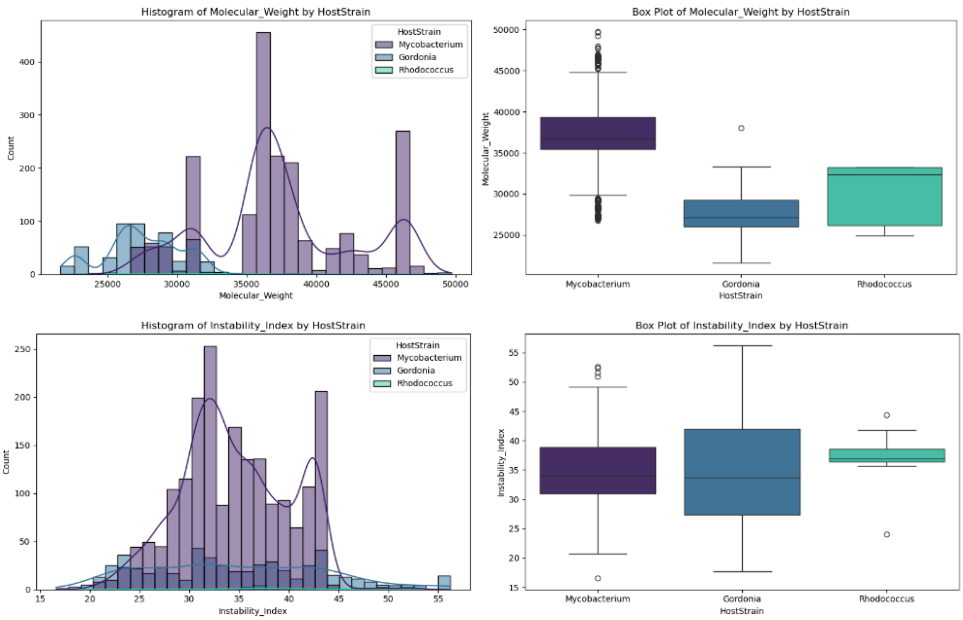
Exploratory Data Analysis And Initial Findings

Visualization of Biochemical Properties by Host Strain

> We aimed to explore the distribution and variability of key features of Lysin B proteins across different Host Strains. We generated histograms and boxplots which allowed us to visualize their distribution and central tendencies, grouped by Host Strain.
>
> For each feature, a histogram was created to show the distribution of values, with a Kernel Density Estimate (KDE) overlaid to highlight the probability density. The distribution was broken down by Host Strain using color differentiation.
>
> Box plots were also created to observe the variability of each feature across different host strains. These visualizations allowed for the comparison of the

> spread and potential outliers in the properties of Lysin B proteins.
>
> Visual Insights:

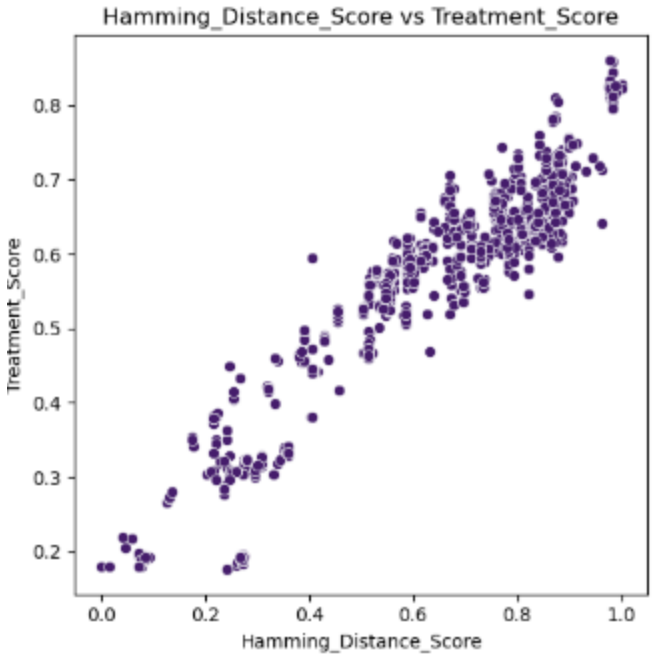
The combination of histograms and boxplots revealed how certain biochemical features varied between host strains. For example, features like Molecular Weight exhibited significant variability, while others like Instability Index appeared more uniform across host strains.

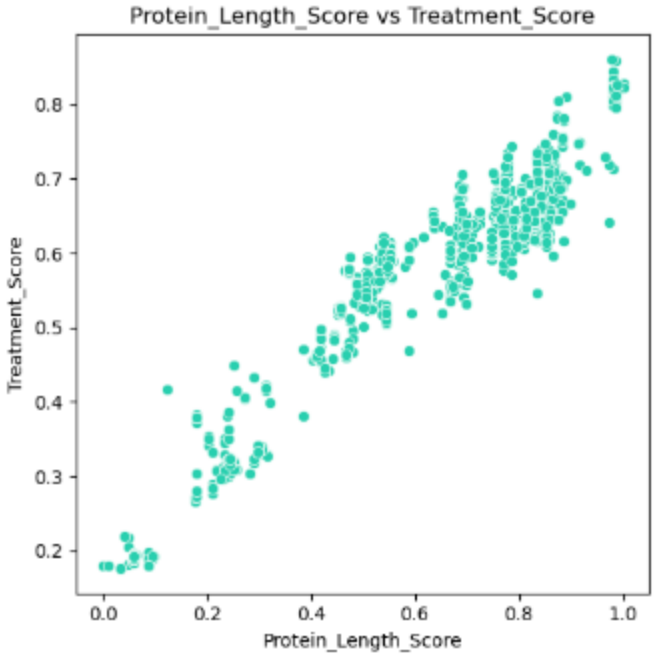


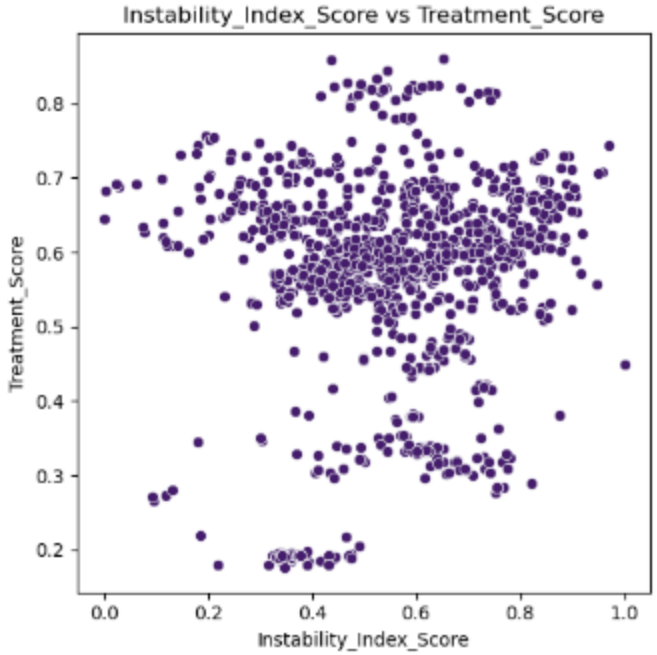
Feature Correlation with Treatment Score

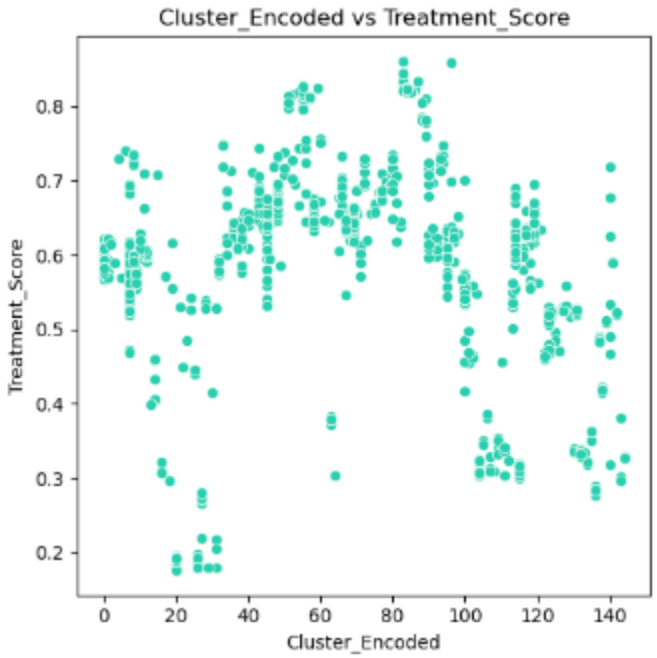
> We examined how various scaled and encoded features correlated with the Treatment Score of each Lysin B protein. The Treatment Score was a calculated value representing the potential therapeutic effectiveness of the Lysin B protein. For each feature, we created scatter plots to visualize their relationship with the Treatment Score. These scatter plots helped determine whether a linear or non-linear relationship existed between each feature and the treatment score.
>
> Visual Insights:

The scatter plots provided a visual representation of how different features influenced the Treatment Score. Features such as Hamming Distance Score and Protein Length Score appeared to have a strong positive correlation with the Treatment Score, while others like Isoelectric Point Score and Cluster Encoded showed a weaker or more complex relationship.


Hamming_Distance_Score vs Treatment_Score
Treatment_Score
Hamming_Distance_Score
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Protein_Length_Score vs Treatment_Score
Treatment_Score
Protein_Length_Score
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8





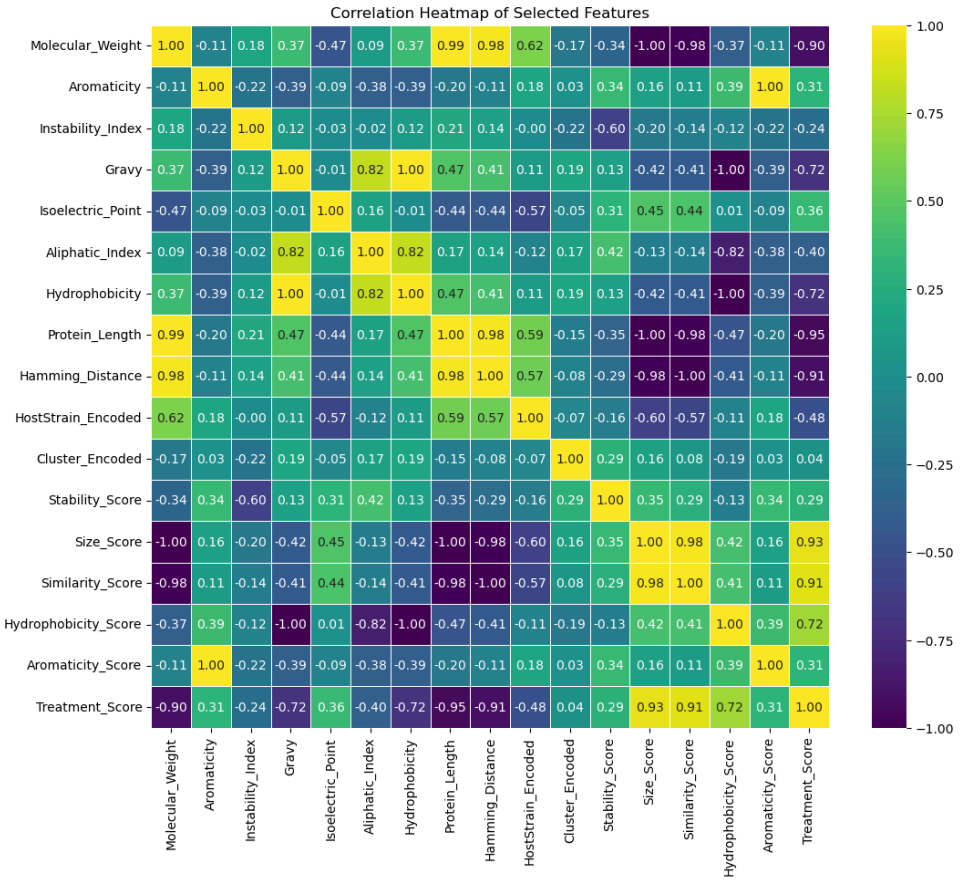
Correlation Analysis

We then focused on quantifying the relationships between various features and the Treatment Score through correlation analysis. We computed the Pearson correlation coefficients between all features using a correlation matrix, which allowed us to identify the strength and direction of relationships. The results were visualized using a heatmap.

Additionally, we extracted and ranked the correlations of each feature with the Treatment Score to determine which features were most closely associated with protein effectiveness.

> Key Findings:

The correlation heatmap revealed key relationships between features. For example, Hydrophobicity Score showed a strong positive correlation, while Protein Length showed a strong negative correlation, indicating their importance in determining Treatment Score. Meanwhile, Cluster Encoded and Instability Index have a very week correlation. These insights were essential for guiding further feature selection and model building.



Model Preprocessing

Feature Selection

> The features relevant to predicting the Treatment Score were defined. The selected features included various biochemical and sequence conservation properties, allowing for comprehensive modeling of the data.

Data Splitting

> The dataset was divided into training and testing sets using an 80-20 split. This separation ensured that the model could be trained on one subset while being evaluated on an unseen subset to assess its performance.

Feature Scaling

> To standardize the data and enhance the performance of machine learning algorithms, the features were scaled using the RobustScaler. This scaler was chosen to mitigate the effects of outliers, providing a more reliable transformation of the data.

Model Selection and Training

> Multiple machine learning models were evaluated to determine which provided the best predictions for the Treatment Score

Model Implementation

> The following models were implemented and trained:
>
> - Linear Regression
> - Ridge Regression
> - Random Forest Regressor
> - Gradient Boosting Regressor
> - Support Vector Regression (SVR)

Each model was trained on the scaled training data, and predictions were made for both the training and test sets.

Model Evaluation

> To assess model performance, a variety of evaluation metrics were calculated, including:
>
> - Mean Squared Error (MSE)
> - Mean Absolute Error (MAE)
> - Root Mean Squared Error (RMSE)
> - R-squared (R^2)
> - Explained Variance
> - Mean Absolute Percentage Error (MAPE)
>
> The results indicated that both the Gradient Boosting and Random Forest models performed well, with the Gradient Boosting model exhibiting the lowest error metrics and the highest R^2 scores.

Hyperparameter Tuning

> GridSearchCV was employed to optimize hyperparameters for both the Random Forest and Gradient Boosting models. This step was crucial in identifying the best parameters that minimized the mean squared error.

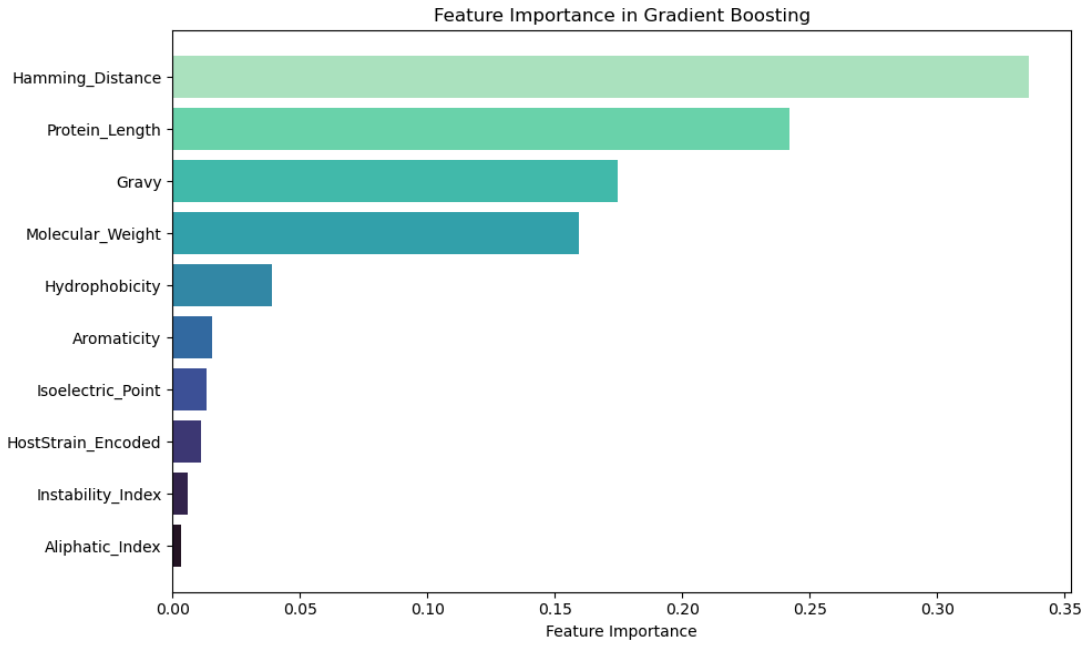
Model	Train MSE	Test MSE	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2	Train Explained Variance	Test Explained Variance	Train MAPE	Test MAPE
Random Forest Grid Search	0.000006	0.000045	0.000955	0.002759	0.002503	0.006726	0.999741	0.997973	0.999741	0.997976	0.001788	0.005271
Gradient Boosting Grid Search	0.000003	0.000031	0.001013	0.002619	0.001650	0.005582	0.999888	0.998604	0.999888	0.998605	0.002012	0.004977

> The Gradient Boosting model yielded the best metrics after performing GridSearchCV and Bayesian Optimization was used to further fine-tune the model. This method enhanced the model's performance, yielding a more accurate prediction of the treatment score.

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 Analysis

> The tuned Gradient Booting model's feature importances were analyzed to identify which features had the most significant impact on predicting the treatment score. A bar chart was generated to visualize the importance of each feature, highlighting those that contributed most to the model's predictive power.



Final Model

> The final Gradient Boosting model was retrained and Treatment Scores were predicted for the entire dataset.

Conclusion

Model Performance and Predictive Accuracy

> Through the course of this project, several significant observations were made that provide critical insights into the performance of lysin-based phage enzyme treatments. Below are the key findings:
>
> - The Gradient Boosting Regressor, after being tuned through Bayesian Optimization, emerged as the best-performing model with the lowest error metrics (MSE, RMSE, MAE) and the highest R^2 and Explained Variance scores.
> - This model demonstrated strong generalization capabilities, meaning it was able to maintain performance consistency across both training and test sets, suggesting that the model is not overfitted and is capable of performing well on unseen data.
> - The final Gradient Boosting model's Mean Absolute Percentage Error (MAPE) on both the training and test sets was low, which indicates that the model consistently provides predictions that are very close to the actual values of the Treatment Score.
> - The model identified lysin proteins with the highest predicted Treatment Scores. Notably, proteins related to Mycobacterium strains (targeting skin and lung infections) showed promising predicted scores. This insight could be pivotal in developing targeted treatments for infections like tuberculosis or Mycobacterium abscessus, which are notoriously difficult to treat with conventional antibiotics.
> - The results from the best-performing model demonstrate that lysin proteins' effectiveness, as measured by the Treatment Score, can be reasonably predicted based on the physicochemical properties of the proteins. This model, therefore, provides a valuable tool for screening potential lysins before conducting costly and time-consuming biological experiments.

Top Predicted Treatment Scores

> After training the final model, we applied it to the entire dataset to predict the Treatment Scores for all lysin proteins. These scores are an indication of how effective each lysin protein is expected to be, not only in phage therapy but also in terms of their biochemical properties relevant for purification and similarity to a consensus sequence. These properties are critical for both practical applications, like drug formulation, and for the generalizability of the treatment to a broader range of bacterial strains.

The predicted Treatment Scores ranged from 0.177 to 0.859. This range reflects the variability in the physicochemical properties of the lysin proteins and their alignment with a biochemical consensus, which helps ensure that these proteins can be effectively purified and maintain consistent efficacy across different bacterial strains.

- Lower Bound (0.177): Proteins with scores near the lower bound likely have characteristics that make them less favorable for both therapeutic efficacy and purification processes. These proteins may exhibit poor structural stability, making them harder to purify, or deviate from the biochemical consensus, reducing their generalizability.
- Upper Bound (0.859): Proteins with scores near the upper bound are predicted to be highly effective not only in treatment but also in practical applications, such as drug purification. These proteins tend to exhibit optimal biochemical properties, such as stable molecular weight and favorable hydrophobicity, making them easier to purify. Additionally, their similarity to a consensus sequence suggests they may work across a broader range of bacterial strains, enhancing their generalizability in real-world applications.

	GeneID	HostStrain	Predicted_Treatment_Score	Treatment_Score
1356	Malisha_CDS_30	Gordonia	0.859357	0.859901
1354	Malibo_CDS_27	Gordonia	0.856586	0.859418
654	Ecliptus_CDS_31	Gordonia	0.843816	0.843529
1204	Kuwabara_CDS_27	Gordonia	0.832911	0.832820
256	Birdsong_CDS_27	Gordonia	0.832911	0.832820
207	BearBQ_CDS_27	Gordonia	0.832911	0.832820
1616	Nymphadora_CDS_25	Gordonia	0.828176	0.828889
1828	Polly_CDS_24	Gordonia	0.825838	0.823357
233	Bialota_CDS_25	Gordonia	0.825624	0.825594
2494	Zirinka_CDS_25	Gordonia	0.825624	0.825594

Focusing on Mycobacterium Strains

Given the project's focus on Mycobacterium infections, specific attention was paid to lysins that target Mycobacterium strains. These strains are of particular interest due to their role in diseases like tuberculosis and Mycobacterium abscessus, which are notoriously difficult to treat with conventional antibiotics.

From the dataset, we filtered out lysins that were designed to target Mycobacterium strains and ranked them according to their predicted Treatment Scores. The top-performing Mycobacterium-targeting lysins

showed Treatment Scores in the higher range of the model's predictions, indicating that they may be particularly effective against these strains. This is promising for the development of therapeutic interventions aimed at Mycobacterium-based infections, which are often resistant to traditional antibiotics.

	GeneID	HostStrain	Predicted_Treatment_Score	Treatment_Score
1167	Kimona_CDS_5	Mycobacterium	0.740391	0.740749
1860	Puppy_CDS_9	Mycobacterium	0.734995	0.734895
1809	Pistachio_CDS_9	Mycobacterium	0.734995	0.734895
1004	Idleandcovert_CDS_8	Mycobacterium	0.734995	0.734895
2265	TNguyen7_CDS_9	Mycobacterium	0.734995	0.734895
274	BlueBird_CDS_9	Mycobacterium	0.734995	0.734895
783	Fred313_CDS_8	Mycobacterium	0.731060	0.731949
164	BabyRay_CDS_9	Mycobacterium	0.729006	0.728869
1327	MA5_CDS_9	Mycobacterium	0.727330	0.726801
1731	Phantastic_CDS_9	Mycobacterium	0.725917	0.727489

Proteins with higher predicted Treatment Scores, particularly those targeting Mycobacterium, should be prioritized for further study and validation. These proteins may offer new therapeutic avenues, especially for treating drug-resistant bacterial infections like those caused by Mycobacterium tuberculosis. The predicted Treatment Scores provide a quantitative measure that can be used to rank and select the best candidates for laboratory testing.

The proteins with the highest predicted scores likely possess features that make them more effective in bacterial cell wall degradation. For example, proteins with optimal isoelectric points and hydrophobicity may have a better ability to interact with and disrupt bacterial membranes, while protein length and molecular weight may contribute to their structural stability and functionality in hostile biological environments.

By using these predictions, researchers can narrow down their focus to the most promising candidates, saving both time and resources in the drug development pipeline.

Future Directions

Incorporating Additional Features

> While physicochemical properties are valuable predictors, integrating additional biological features could enhance the model's accuracy. These could include:
>
> - Protein structural data (e.g., secondary or tertiary structure, domain-specific properties)
> - Post-translational modifications that might affect protein stability or interaction with bacterial membranes.
> - Binding affinities between lysins and bacterial cell walls.

Experimental Validation

> The ultimate goal of this project is to validate the model's predictions in a laboratory setting by testing the top-predicted phage lysins on bacterial cultures. This will help determine if the predicted Treatment Scores—which estimate how effective a lysin protein is at breaking down bacterial cell walls—match the actual biological results. The primary objective would be to see how well these lysins perform against harmful bacterial strains, particularly those that are difficult to treat, such as Mycobacterium.
>
> Before moving on to testing in bacterial cultures, it is also important to assess the ease of purification of the top-predicted lysins. The model has identified lysins with favorable biochemical properties, such as optimal stability, molecular weight, and isoelectric point, which influence how easy it is to purify these proteins. Purification is a critical step because, to be used in real-world applications like drug development, lysins must be produced in large quantities. Lysins that are easier to purify and manufacture are more likely to be practical for therapeutic use.
>
> Additionally, the lab testing phase would include experimenting with combination treatments. Lysins could be tested alongside other therapies, such as phage therapy or antibiotics, to see if using them together makes treatment more effective. Phage lysins might enhance the ability of bacteriophages (viruses that infect and kill bacteria) to degrade bacterial cell walls. When combined with antibiotics, lysins may help reduce the amount of antibiotics needed or make the treatment more effective against antibiotic-resistant bacteria.
>
> By combining computational predictions with experimental validation, this project provides a foundation for identifying lysins that are not only effective on their own but also enhance other treatments, such as phages or antibiotics. This approach could lead to more scalable and potent therapeutic options.

In [ ]: