Bacterial Virulence Prediction: Group 22

Hannah Bower, Sanya Badole, Kristin Keith, and Rabab Fatma

I. Abstract

Accurate identification of virulent bacterial strains is essential for furthering infectious disease research and treatment. Current methods remain dependent on experimental assays that are time-consuming and resource-intensive. Our pipeline outlines a computational framework to predict virulence across diverse bacterial species by analyzing protein folding patterns using AlphaFold 2, a high performing 3D protein structure prediction model. This approach enables high-resolution analysis of 3D protein structures to identify mechanistic drivers of pathogenicity. By using large-scale protein structure predictions, we are aiming to find patterns in 3D protein shapes that are linked to how harmful or virulent the bacteria can be. This approach enables efficient screening of large protein sets and supports the discovery of structural markers that may underlie pathogenicity, host interaction, and evolutionary adaptation. Our pipeline could inform the development of novel antimicrobial targets and improve understanding of the molecular mechanisms driving bacterial virulence.

On our initial dataset, we achieved an AUC of 0.94 using logistic regression trained on structure-based embeddings. In contrast, models trained only on sequence-based embeddings performed with an AUC of 0.5, which highlights how much stronger structural feature prediction is. We further improved our model by tuning the hyperparameters and building an ensemble classifier (Voting Classifier) that combines the Random Forest, XGBoost, CatBoost and Light GBM models. This optimized pipeline then reached a validation accuracy of 88% and a test accuracy of 87.4% with strong precision and recall for both virulent and non-virulent classes.

Our results demonstrate that structure-based embeddings derived from AlphaFold 2 provide significant predictive power for bacterial protein virulence, enabling rapid and scalable screening of large protein datasets. The structural focus has the potential to help accelerate the identification of new virulence factors as well as find new targets for antibacterial drugs and most of all deepen the understanding of what causes pathogenicity in bacteria.

II. Introduction

Bacteria are one of the most prevalent organisms on earth, ranging from being beneficial symbionts to deadly disease-causing pathogens [1]. Virulence in a bacterium is a way of measuring the likelihood of causing disease. Virulence is determined by a variety of factors that include toxins, adhesion molecules, immune evasion proteins and secretion systems. Databases like the Virulence Factor Database (VFDB) have cataloged thousands of these known bacterial virulence factors across diverse species of bacterium. Factors like gene presence and absence, genetic variation and regulatory differences contribute to the variability in pathogenicity [2]. Being able to accurately identify virulent strains and genes is extremely important in public health monitoring as well as treatment in a clinical setting. Traditional identification methods that phenotype one organism at a time are extremely labor intensive, and machine learning on whole genome sequences can overlook specific important traits [3].

Advances in protein structure prediction with models like AlphaFold allow the creation of highly accurate 3D structures just from amino acid sequences without sequence alignment [4]. This tool specifically introduces a potential for large scale analysis of the structure of bacterial proteomes. The structure of a protein plays a huge role in the determination of function, and studying them could show important patterns in virulence and non virulence. We are hypothesizing that bacterial virulence is associated with

specific protein folding patterns that distinguish virulence factors from non virulence proteins within the proteome. We introduce a pipeline that combines AlphaFold structural predictions with sequence embeddings to train a machine learning model to predict pathogenicity in bacterial proteomes. By focusing on structural indicators of virulence, we hope to contribute to new therapeutic targets to help treat bacterial infections and most importantly the ones that are more virulent.

III. Related Work

Prediction of bacterial virulence has been a popular subject in recent years, but our pipeline using protein structural data to predict bacterial virulence is a novel approach. One published work by Quintana et al. is very closely related to our approach, except it uses a two stage model approach with a neural network and a GCN [5]. "VF-Pred: Predicting virulence factor using sequence alignment percentage and ensemble learning models" [6] is another piece of work that is closely related to our pipeline. However, this work primarily focuses on sequence based features, not including structural data as our pipeline does. A study by Wang et al. used whole genome sequencing and machine learning to analyze *S. aureus* strains [7]. However, the researchers focused primarily on genetic traits associated with antimicrobial resistance rather than using our approach with protein structure. A study done by Li et al. utilized ESMFold rather than AlphaFold for their structure predictions. They also used a graph based representation rather than our structural embeddings approach [8]. Additionally, AlphaFold predictions have shown promise with accurate protein structure prediction, as it is being used by Miller et al. for protein function based on structure[9]. This paper is promising as it supports our idea that protein structure may be able to help indicate virulence.

All of these related studies show growing interest in new approaches to bacterial pathogenicity. Currently, none have combined high-resolution protein modeling with machine learning to predict virulence in bacterial proteomes. In our pipeline that focuses on 3D protein modeling, we introduce new biological insights that complement current sequence-based methods. This pipeline may highlight structural patterns that are directly tied to virulence.

IV. Methods

We began by curating protein-level data specific to bacteria from the Virulence Factor Database (VFDB) that were categorized as virulent or non-virulent. Virulent protein sequences and their RefSeqIDs were downloaded from VFDB, converted to UniProtID via UnitProt's API, and their corresponding 3D structures were retrieved from the AlphaFoldDB[10]. Similarly, non-virulent proteins were collected from

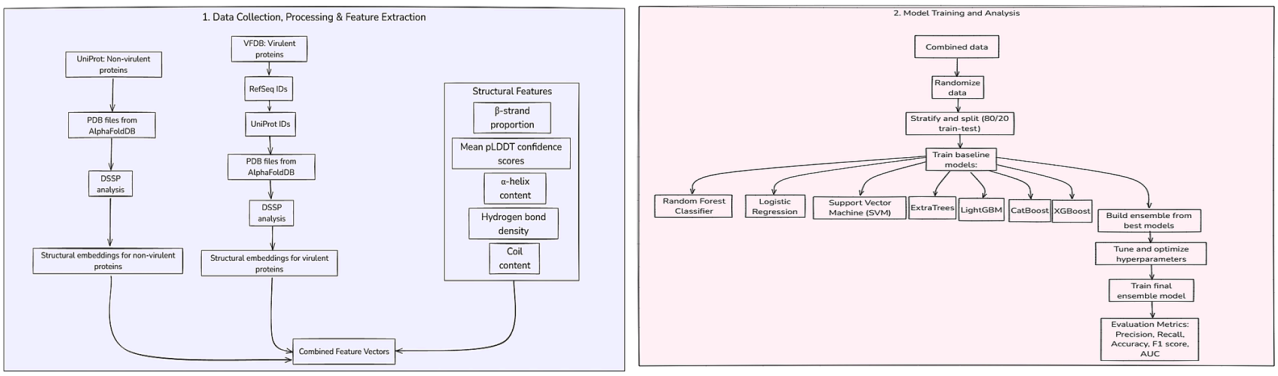


Figure 1. Workflow charts for data collection and processing (left) and model training (right)

UniProt, including proteins from non-pathogenic species for class imbalance. Structural features such as alpha helix, beta sheet, and coil content, as well as hydrogen bond density and mean pLDDT were extracted using DSSP and compiled into two CSV files. These files contained approximately 15,000 virulent and 8,000 non-virulent proteins.

Assuring data quality and eliminating artifacts that can affect model performance were our top priorities during preprocessing. Proteins lacking any of the essential structural characteristics (NaNs) were removed. Incomplete feature vectors can distort model training and produce incorrect results, therefore this filtering phase was essential. The virulent and non-virulent datasets were concatenated into a single, cohesive dataset following cleaning with the addition of a label column that contained binary values (0 for non-virulent and 1 for virulent). We ensure reproducibility of the training-testing split by randomly selecting a fixed random seed (42), which prevents any ordering artifacts from distorting the learning process.

We utilized structure-based features to convert the data into usable inputs for our models. To predict protein structure, we used AlphaFold2 which was implemented through ColabFold, for our ease of use [4, 8]. Initially, we considered ESMFold as a prediction software due to its faster runtime, but installation challenges led us to choose the AlphaFold2 server on Google Colab. By outsourcing our structure predictions, we streamline our processes by avoiding configuring AlphaFold locally and are able to achieve efficient and accurate results. We chose to go with the default options on the ColabFold platform parameters except for the recycling reiterations in which we chose 6. In a study by Montiero et al., the researchers found that by doubling the amount of recycling steps in AlphaFold the model becomes much more confident and often predicts the most stable structure and decreases the amount of less certain results, which is why we chose to double the amount of recycles from the default amount of 3 [9].

We implemented Biopython, pandas, Torch, and the ESM model suite to extract multiple biologically relevant features from our predicted proteins. DSSP secondary structure assignments were again used to find the proportion of β-strand residues to predict β-sheet content, and mean pLDDT confidence scores were calculated from B-factor fields of AlphaFold PDB files. Alpha-helix content, hydrogen bond density, and coil content were also extracted from the data. ESM model loading was implemented outside per-protein loops to speed up processing rate. A final CSV file containing these vectors was generated for use in modeling steps.

From our final CSV file, we eliminated identifiers like UniProt IDs and only used numerical features that were physically significant. Alpha-helix content, beta-sheet content, coil content, hydrogen bond density, and mean pLDDT scores made up feature matrix X. The label vector y was made up of binary virulence indicators. The percentage of virulent and non-virulent samples in each subset was then maintained by doing an 70/15/15 stratified train-validation-test split. Because bacterial virulence variables were not equally distributed between the two classes (due to the data imbalance), stratification was essential, and preserving class balance was required to guarantee fair evaluation.

We chose classical machine learning models for our initial predictive modeling. Our model was trained on the following classifiers: Random Forest Classifier, Logistic Regression, Support Vector Machine (SVM), ExtraTrees, LightGBM, CatBoost, and XGBoost. We expected that modeling structural determinants of virulence would require the ability to capture complex, non-linear feature interactions, which is why tree-based ensemble methods like Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost were chosen. To start, we used default model parameters to obtain a performance baseline and saved hyperparameter tuning for downstream work. Our models were evaluated on precision, recall, accuracy,

F1 score, and ROC-AUC metrics. Initial results showed that logistic regression reached 88.24% accuracy and tree-based models achieved around 78-80% accuracy.

After assessing each model separately, we found that although a number of classifiers performed well, no single model consistently outperformed the others on all metrics. We chose to create an ensemble model after realizing the inherent biases and limits of individual models, such as decision trees' propensity to overfit or logistic regression's sensitivity to feature scaling. Combining the probabilistic results of the Random Forest, XGBoost, CatBoost, and LightGBM models, the ensemble used a soft voting technique. This strategy was used because, when individual classifiers are as diverse as ours were, ensembles tend to generalize better than single models by lowering model-specific variance and bias.

Finally, to further enhance our ensemble model performance, we implemented grid search (to tune n_estimators in tree-based models) and Bayesian optimization (to tune more sensitive parameters such as max_depth, iterations, and learning_rate). Using Bayesian optimization improved validation accuracy from 87.04% to 87.76% and test accuracy from 86.04% to 87.10%. This proves that hyperparameter tuning was essential in unlocking the full potential of using tree-based methods in our training.

V. Data

The structural protein-level data used in this study came from a variety of bacterial species, and each protein was classified as either virulent or non-virulent. The Virulence Factors Database (VFDB) , a carefully selected and scientifically verified collection of bacterial virulence factors from a broad range of taxa, provided the virulent protein sequences [2]. We obtained 25,141 RefSeq identifiers for the virulence -associated proteins from VFDB. These were then converted to UniPort ID’s for compatibility with tools used in downstream analysis. We initially tried using UniProt’s API, but querying over 25,000 sequence IDs was taking a significant amount of time. Instead, we downloaded the Uniprot mapping file which was 11GB in total and used grep commands to map the RefSeq IDs to UniProt IDs.

Non-virulent proteins were obtained from UniProt by filtering out keywords associated with virulence. The number of non-virulent proteins was found to be less than virulent associated proteins as unlike VFDB, there is no specific database for the collection of non-virulent proteins; hence our collection was limited to well annotated UniProt entries. To somewhat counteract this data imbalance we decided to include data from non-pathogenic bacterial species such as *Bacillus subtilis*, which are not commonly not linked to virulence in order to increase the representation of the negative class. Once both non-virulent and virulent UniProt ID’s were collected, we retrieved precomputed 3D structures from the AlphaFold Protein Structure Database using these identifiers. The structural foundation of this investigation was selected to be AlphaFold2, a deep learning-based protein structure prediction framework created by DeepMind that is well known for its remarkable accuracy [12].

After the structural models were retrieved in PDB (Protein Data Bank) format, we used DSSP (Dictionary of Secondary Structure of Proteins) to extract features. DSSP is the standard algorithm for assigning secondary structure to amino acids in a protein based on atomic coordinates [13]. Five biologically significant structural features were extracted for each protein; the mean pLDDT score, h-bond density, coil content, beta-sheet content and alpha-helix content. To represent the glocal model confidence, we averaged the pLDDT score - a per-residue confidence metric produced by AlphaFold - across the residues.

23,783 proteins made up the final cleaned dataset following all preprocessing steps (around 15,000 virulent and 8,000 non-virulent). Although there was still class disparity, it was far less pronounced that it had been at first. To remove ordering bias the dataset was mixed at random using a fixed seed and this

was divided into a 15% validation and test set each and a 70% training set. 16,783 proteins made up the training set, while 3,596 proteins made up the validation and test splits each. To ensure that virulent and non-virulent proteins were proportionately represented in each subset, stratified sampling was used. Our pipeline's downstream classifying tasks were built upon these carefully selected and annotated structural profiles.
An exploratory data analysis was done and is present at the end as supplementary information.

VI. Experiments

We carried out a number of tests utilizing both individual machine learning models and ensemble approaches to assess how well structural factors predict bacterial pathogenicity. Finding out if high-resolution protein structure data could accurately differentiate between virulent and nonvirulent bacterial proteins was our main objective. Four common classifiers–Random Forest, Logistic Regression, XGBoost and a Multi-Layer Perceptron (MLP) neural network–were used for baseline assessments. Random Forest and XGBoost were chosen for their capacity to model non-linear relationships and their strong performance on structured tabular data [14,15], whereas logistic regression was used as linear baseline to assess whether the data showed linear separability[16].We were able to compare deep learning techniques with conventional ensemble methods thanks to MLP's provision of a neural network benchmark [11]. According to our preliminary findings, tree-based models such as Random Forest and XGBoost routinely outperform logistic regression and neural nets **(Table 1).** This suggests that interpretable, non-linear ensemble techniques, rather than linearly separate or high-capacity deep-learning models, are better suited to capture virulence patterns. Due to our previous finding, we decided to evaluate the baseline performance of other decision tree models, namely CatBoost and LightGBM. These models are advanced gradient boosting frameworks designed for efficiency and accuracy on tabular data, with CatBoost excelling in categorical feature handling and LightGBM optimized for speed and scalability [17,18]. With high accuracy values, this validated our hypothesis that decision trees are the optimal model for our problem.To improve the overall classification performance and lessen the individual shortcomings of standalone models, we adopted a more thorough ensemble-based approach after the baseline study. Using the 4 best performing tree-based models–Random Forest, XGBoost, CatBoost and LightGBM–we built a soft voting ensemble classifier. Using grid search, each of these models was independently adjusted for important hyperparameters such as learning rate, maximum tree depth, number of estimators, and boosting iterations. For example, XGBoost and LightGBM were tweaked for boosting rounds and shrinkage rates, whereas Random Forest was tuned across combinations of estimator count and tree depth. To prevent overfitting, grid search was performed on the training set using 3-fold cross-validation.

Model	Accuracy
Random Forest	0.87
Logistic Regression	0.76
XGBoost	0.85
MLP Neural Net	0.73
Support Vector Machine (SVM)	0.63
K-Nearest Neighbours (KNN)	0.75
CatBoost	0.80
LightGBM	0.80

Table 1 - Baseline accuracy performance of various ML models

Following optimization, a soft voting technique was used to

merge the various models, with the final prediction being determined by averaging the projected probabilities among classifiers. This method enabled the ensembl to take advantage of complement strengths. With a balanced performance between precision and recall for both virulent and non-virulent classes, the final ensemble had an accuracy of 88% on the validation set. The model maintained a high accuracy of 87.4% when tested on the independent test set, indicating significant generalization (Table 2). Excellent discriminative capability was indicated by the ROC curve's (Figure 3) AUC of 0.94.

Model	Validation Accuracy	Test Accuracy	Validation F1	Test F1	Validation ROC AUC	Test ROC AUC
Random Forest	0.875	0.871	0.873	0.8695	0.941	0.935
XGBoost	0.859	0.846	0.857	0.844	0.921	0.9151
CatBoost	0.846	0.839	0.844	0.836	0.913	0.906
LightGBM	0.847	0.838	0.844	0.835	0.911	0.904
KNN	0.775	0.759	0.770	0.753	0.812	0.802
MLP Neural Net	0.758	0.7484	0.747	0.737	0.819	0.814
SVM (RBF Kernel)	0.645	0.6452	0.506	0.506	0.724	0.712
Super Voting Ensemble	0.880	0.874	0.880	0.873	0.945	0.940

Table 2 - Performance metrics of various machine learning models after hyperparameter optimization. Metrics include validation and test accuracy, F1 score, and ROC AUC. The Super Voting Ensemble achieves the highest overall performance, with a test ROC AUC of 0.94, followed closely by Random Forest. Tree-based models generally outperform neural networks and SVM, confirming their suitability for this bacterial pathogenicity prediction task.

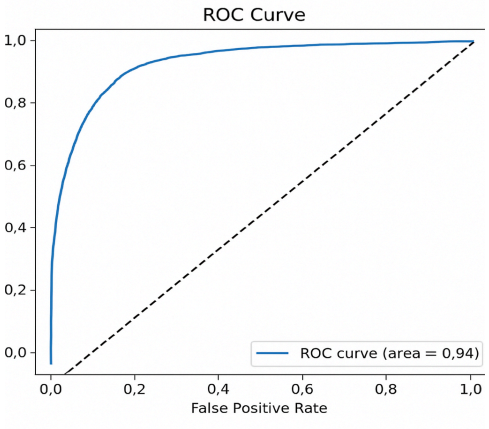


Figure 3 - ROC AUC Graph of Super Voting Ensemble

We carried out an ablation study by gradually eliminating each feature from the dataset and tracking the resulting decline in validation accuracy in order to better understand which characteristics contributed the most to the model's performance. The final ensemble model served as the foundation classifier for this analysis. The two most important characteristics, according to the results, were mean pLDDT and hydrogen-bond density. A notable accuracy decrease of about 12% and 10% respectively, resulted from removing any of them (Figure 4)**.** On the other hand, when left out, characteristics like coil and alpha helix content exhibited only slight effects. The results emphasize how crucial structural stability (as detected by pLDDT) and interaction patterns (demonstrated by hydrogen bonding) are in differentiation protein behaviour linked to virulence.

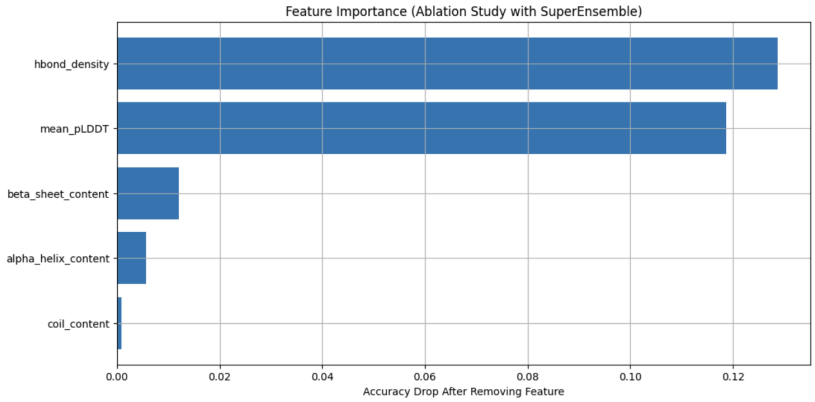


Figure 4 - Feature importance from an ablation study using the SuperEnsemble model. The horizontal bar chart displays the impact of removing each feature on model accuracy, with hbond_density and mean_pLDDT having the greatest influence on performance, while features like coil_content contribute minimally.

We used both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) to project the high-dimensional feature space into 2 dimensions for additional visualization (Figure 5). The virulent and nonvirulent proteins formed partially separable clusters with distinct but on-linear borders, according to these projections. This confirmed our previous finding that tree-based approaches are more suited for this task than linear models.

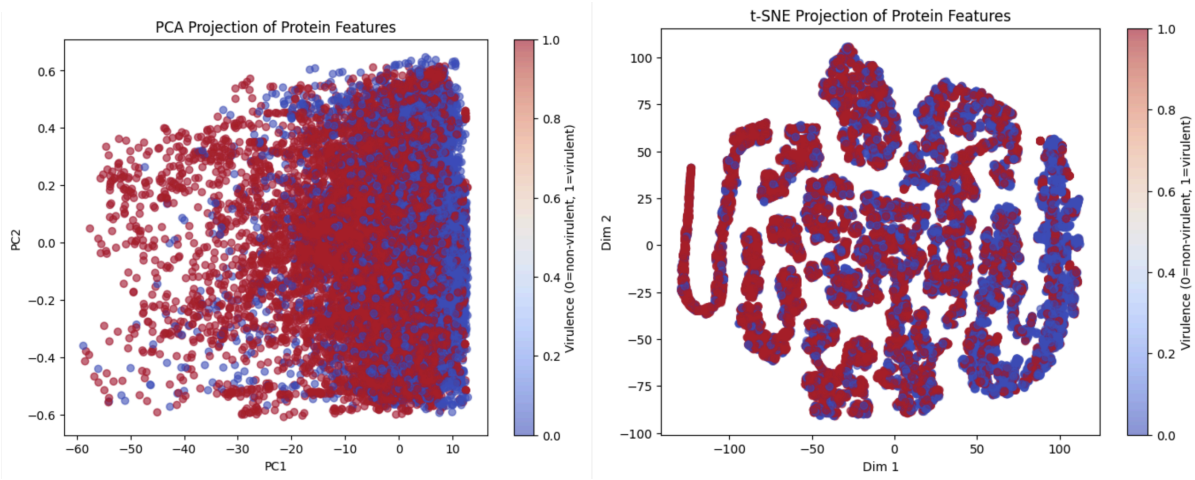


Figure 5.PCA (left) and t-SNE (right) projections of structural protein features for virulent (redd) and non-virulent(blue) bacterial proteins.

The majority of classification includes proteins with low mean pLDDT scores, indicating that classification accuracy may be impacted by confidence in structure prediction, even though the model performed well overall. The flexible or disordered areas as seen in these low-confidence proteins may make it more difficult for AlphaFold to represent them accurately and could contribute nouse into the computation of structural features.

By applying the same models to protein sequences obtained from UniProt IDs, we were also able to assess baseline categorization performance using sequence embeddings produced by ProtBERT. With an

area under the ROC curve (AUC) of 0.94, the structural feature-based model significantly outperformed the ProtBERT sequence embedding-based model in terms of discriminative capacity. This significant difference suggests that structural characteristics obtained from high-confidence AlphaFold2 models have a significantly higher predictive value than sequence-based embeddings alone in our dataset for differentiating between virulent and non-virulent bacterial proteins. This finding emphasizes how using protein structure information for virulence prediction adds value beyond what sequence features can capture.

Model	AUC	Accuracy
Random Forest	0.5	0.789474
Logistic Regression	0.5	0.789474
SVM	0.5	0.789474
LightGBM	0.5	0.789474
XGBoost	0.5	0.789474
VotingEnsemble	0.5	0.789474

Table 3 - Performance metrics of baseline machine learning models trained solely ons sequence embeddings. All models exhibit an AUC of 0.5 and an accuracy of approximately 0.79, indicating that sequence-only embeddings are insufficient for accurately predicting bacterial virulence.

VII. Conclusion and Discussion

In conclusion, we developed a pipeline that leverages AlphaFold 2 predicted protein structures and structure-based embeddings to predict virulence in bacteria. Our findings show that the predictive power of structure- and sequence-based embeddings are significantly higher than sequence embeddings alone. This is indicated by the AUC achieving 0.94 as compared to ~0.5 for sequence only models. Further improvements through hyperparameter optimization and ensemble modeling allowed us to achieve a validation accuracy of 88% as well as a test accuracy of 87.4% with very strong performance in both virulent and non-virulent classes.

Our work highlights the important role of protein structure in the pathogenicity of bacteria. We also demonstrated that structural modeling can be a powerful predictive tool for bacterial virulence. Compared to previous approaches that rely on sequence information as the primary predictor, our structure-centered framework provides new insights into the molecular mechanisms driving virulence. Most importantly, our approach allows for rapid and scalable screening of large bacterial proteomes and paves the way toward a new way of identification of novel virulence and novel antibacterial drug targets.

However, there are many limitations to our framework. Our model relies on known and predicted virulence factors, meaning that new or uncharacterized virulence mechanisms may be missed. Additionally, while AlphaFold 2 is known for producing highly accurate structural predictions, low confidence regions in the predicted models could introduce noise into feature extraction. Future improvements and work could incorporate this uncertainty in calculations, larger datasets, and a graph based model to increase predictive accuracy. Our dependence on static structural components is another possible drawback. These descriptors are helpful, but they might not adequately represent the dynamic activity of proteins, such as post-translational modifications or structural changes during host interaction, which may also be factors in virulence. Predictive performance could be enhanced and the feature set further enhanced by integration with proteomics-derived flexibility data or molecular dynamics simulations.

Overall, this study shows that structure-based machine learning is both feasible and useful for predicting bacterial pathogenicity. Through the use of ensemble learning and AlphaFold-generated structural data, we were able to identify important structural characteristics linked to virulence and obtain good predictive performance. Future research into pathogenicity mechanisms and therapeutic target discovery may benefit from the insights this framework produces, which could ultimately lead to the creation of precision antibiotics and better diagnostics.

VIII. Contributions

Hannah: Feature Extraction, Literature Reading

Sanya: Preprocessing Data, Model development

Kristin: Data Processing and Optimization

Rabab: Figures

All: Report, Running Colab Fold

IX. Source Code

Our source code is located in the following GitHub Repository:

https://github.com/BowerH/Bacterial_Virulence_Prediction

X. References

1. Soni, Jyoti, Sristi Sinha, and Rajesh Pandey. 2024. “Understanding Bacterial Pathogenicity: A Closer Look at the Journey of Harmful Microbes.” *Frontiers in Microbiology* 15 (February). https://doi.org/10.3389/fmicb.2024.1370818

2. Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2004). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research, 33*, D325 - D328.

3. Stromberg ZR, Phillips S, Omberg KM, Hess BM. High-throughput functional trait testing for bacterial pathogens. mSphere. 2023 Sep 13;8(5).

4. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly Accurate Protein Structure Prediction with Alphafold. Nature. 2021 Jul 15;596(7873):583–9.

5. Quintana, Felix, Todd Treangen, and Lydia Kavraki. 2023. “Leveraging Large Language Models for Predicting Microbial Virulence from Protein Structure and Sequence,” September. https://doi.org/10.1145/3584371.3612953.

6. Singh S, Le, Wang C. VF-Pred: Predicting virulence factor using sequence alignment percentage and ensemble learning models. Computers in Biology and Medicine. 2024 Jan 1;168(0010-4825):107662–2.

7. Li G, Bai P, Chen J, Liang C. Identifying virulence factors using graph transformer autoencoder with ESMFold-predicted structures. Computers in Biology and Medicine. 2024 Jan 30;170:108062–2.

8. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nature Methods [Internet]. 2022 May 30;19(6):1–4. Available from: https://www.nature.com/articles/s41592-022-01488-1

9. Monteiro G, Cui JY, Dalgarno DC, Lisi GP, Rubenstein BM. Predicting Relative Populations of Protein Conformations without a Physics Engine Using AlphaFold2. bioRxiv (Cold Spring Harbor Laboratory) [Internet]. 2023 Jul 27 [cited 2024 May 5]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10402055/

10. AlphaFold. AlphaFold Protein Structure Database [Internet]. alphafold.ebi.ac.uk. 2022. Available from: https://alphafold.ebi.ac.uk/

11. Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)-A review of applications in the atmospheric sciences. Atmospheric Environment, 32(14–15), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

12. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

13. Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12), 2577–2637. https://doi.org/10.1002/bip.360221211

14. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

15. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(pp. 785–794). https://doi.org/10.1145/2939672.2939785

16. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

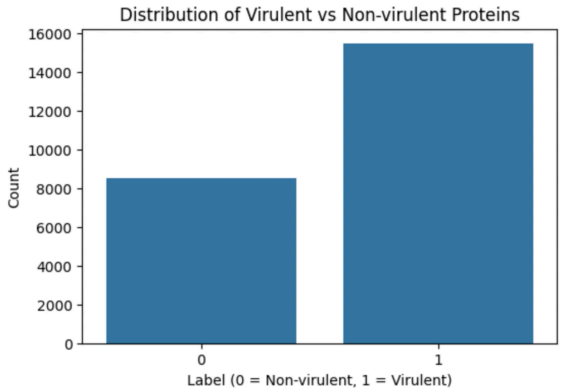
17. Anghel, A., Papandreou, N., Sips, M., & Schumacher, M. (2020). CatBoost for big data: an interdisciplinary review. *Big Data and Cognitive Computing, 4*(4), 1-24. https://doi.org/10.3390/bdcc4040027

18. Amat Rodrigo, J., & Escobar Ortiz, J. (2025). skforecast (Version 0.15.1) [Computer software]. https://doi.org/10.5281/zenodo.8382788

Supplementary Information

1. Exploratory Data Analysis

We carried out a thorough exploratory data analysis to comprehend the dataset's underlying structure. There were roughly 15,000 virulent proteins and 8,000 non-virulent proteins in the final combined dataset. Although there was still a class disparity, it was far less pronounced than it had been at first. Figure 6 displays the class distribution, which confirms a moderate imbalance.



Furthermore, the distribution of AlphaFold's mean pLDDT scores showed that most proteins had a high degree of structural confidence, with scores primarily falling between 85 and 95 . A multimodal distribution of the alpha-helix content was observed, indicating structural variability in the sample.

Different statistical trends were also seen in other structural elements. Most proteins had relatively low beta-sheet proportions, and the number of beta-sheets was strongly skewed to the right. The hydrogen bond density, on the other hand, was concentrated between 0.25 and 0.45, exhibiting a heavy-tailed distribution, whereas coil content followed a peaked distribution centered around 0.4. Complex linkages that were unlikely to be linearly separable were suggested by these irregular patterns across characteristics.

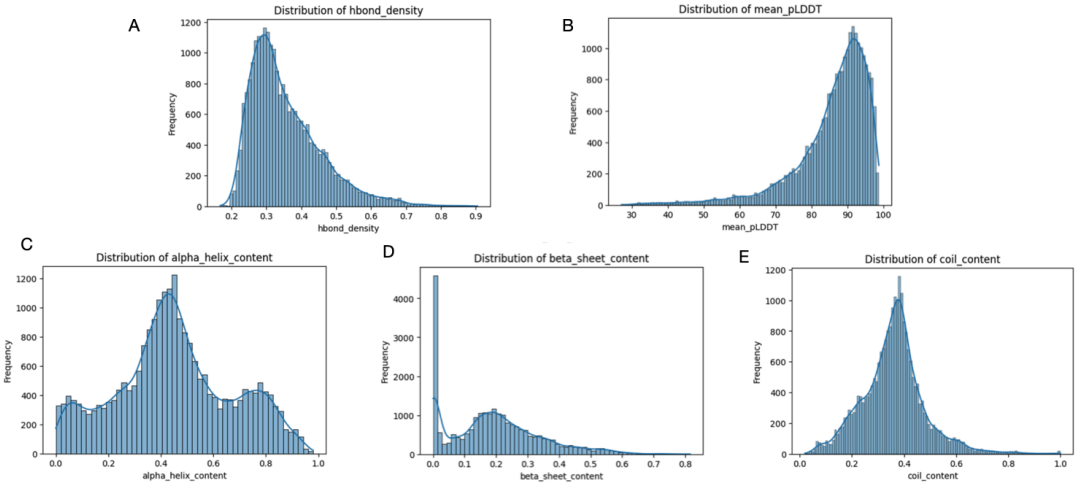


Figure 7 - Distribution of feature values across the dataset

Class-level differences were shown to be minor but constant in this visualization. For instance, the alpha-helix and beta-sheet content of viral proteins was generally higher, whereas the coil content of non-virulent proteins was more widely distributed. The use of nonlinear classifiers was encouraged by the density differences, which suggested learnable structure-function links even if no one feature provided complete class separation.

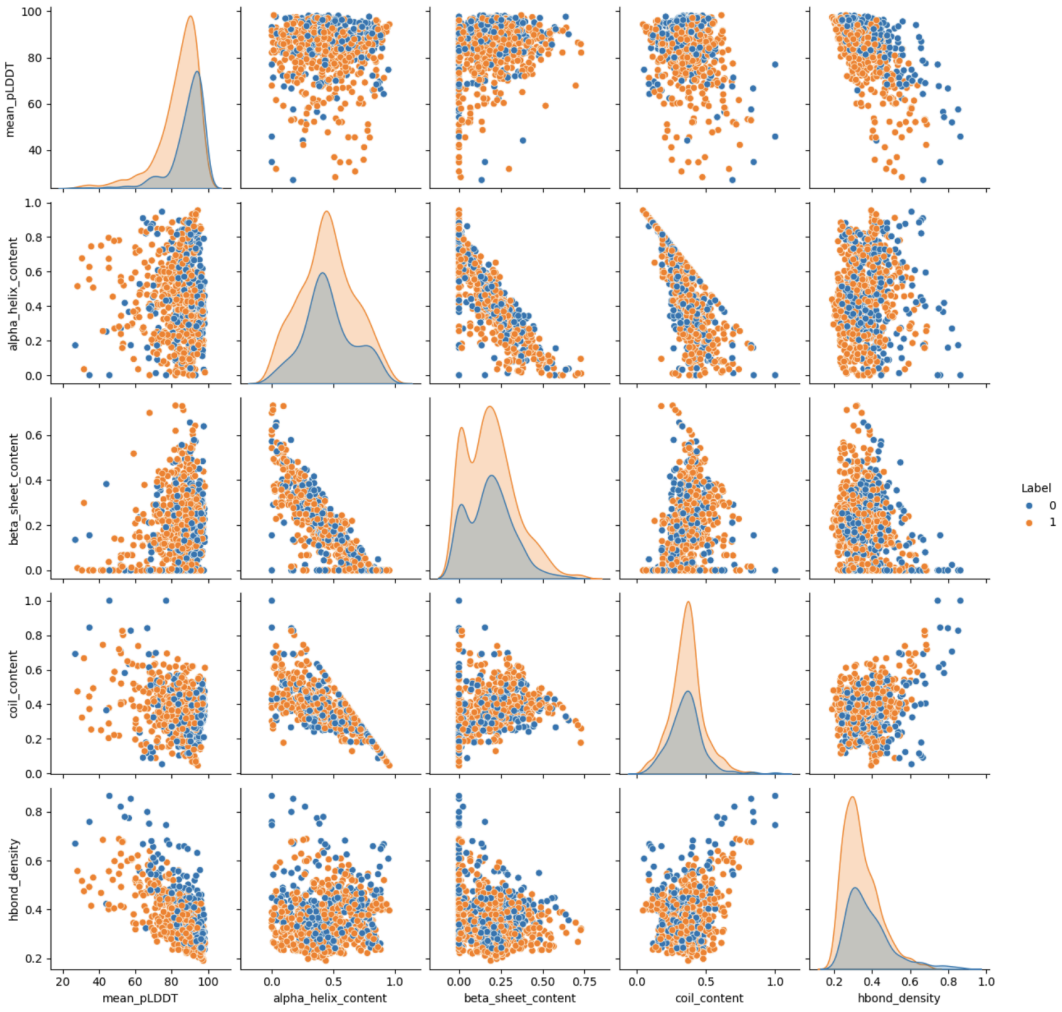


Figure 8 - Pairwise plot showing the relationships and distributions of structural features (mean pLDDT, alpha helix content, beta sheet content, coil content and h-bond density) for virulent (orange) and non virulent (blue) bacterial proteins.

Boxplots offered more proof of these patterns. The distributions of hydrogen bond density, coil content, and beta-sheet content varied somewhat between the virulent and non-virulent classes, as seen in Figure 8. The medians and interquartile spreads displayed slight shifts, despite the ranges' significant overlap, suggesting possible biases in protein folding patterns linked to virulence.

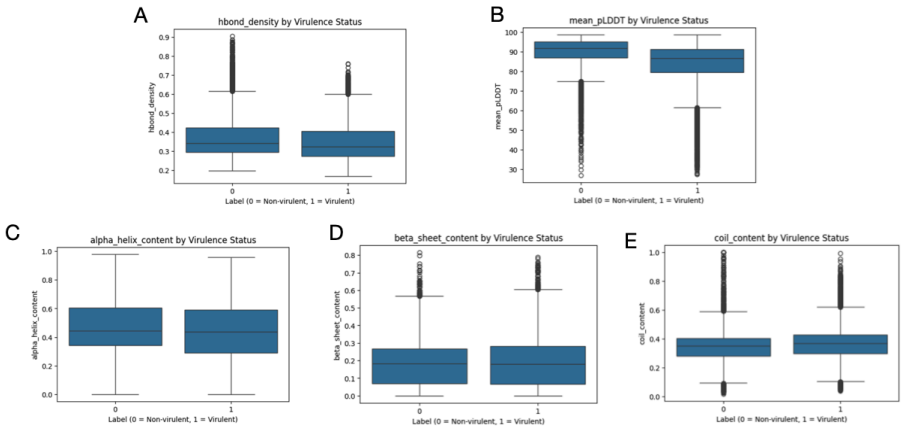


Figure 9 - Box plots showing the distribution of feature values across virulent and non-virulent proteins.

Lastly, a correlation heatmap showed that most features had low to moderate connection with one another. As anticipated, because alpha-helix and coil content complement one another in secondary structure assignment, they had a weakly negative correlation. Machine learning algorithms were able to learn a variety of non-redundant signal patterns from the data because of the generally low levels of multicollinearity, which was advantageous for model training.

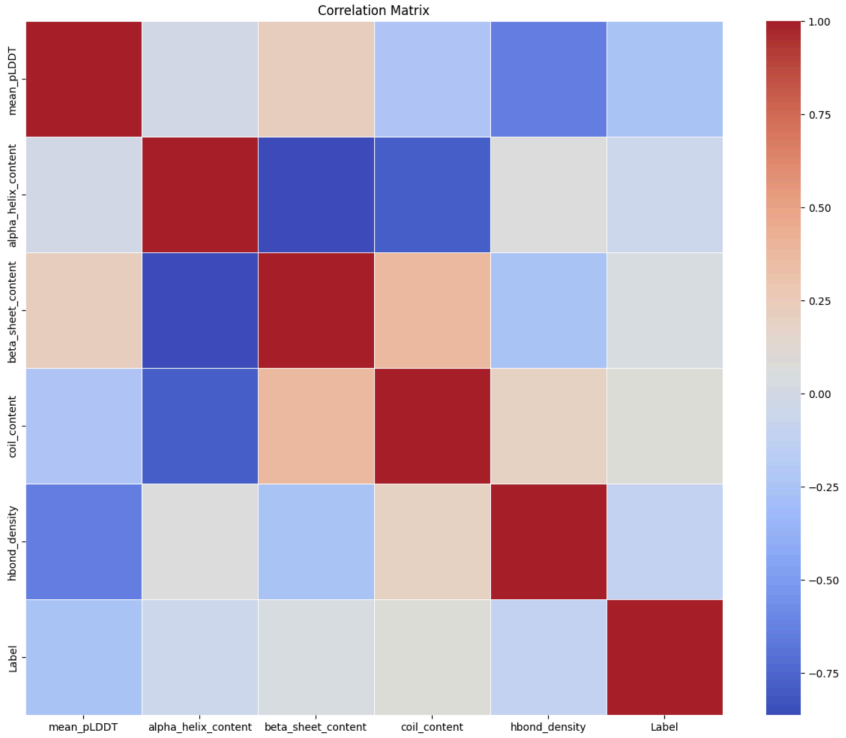


Figure 10 - Correlation heatmap between structural features of virulent and non-virulent proteins used in model development and training.