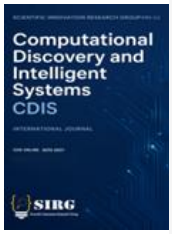

Computational Discovery and Intelligent Systems
CDIS

Journal Homepage
https://pub.scientificirg.com/index.php/CDIS/index

Leveraging Artificial Intelligence for Protein-Based Drug Target Prediction in Pseudomonas aeruginosa

Rajdeep Chakraborty[a,1], Yosra Diaa[b], Ganna Mahmoud[c], Jumana Mohamed[d], Rehab Ahmed[e], Sara Ramadan[f]

[a]Dept. Of CSE, SAGE University, Indore, India, Email: rajdeep_chak@yahoo.co.in
[,b,c,d,e,f]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City,62511, Egypt.
Emails: yousradiaa1067@fcis.bsu.edu.eg , gannamahmoud98_sd@fcis.bsu.edu.eg, jumanaahmed97_sd@fcis.bsu.edu.eg , rehabahmed23234@gmail.com , sararamadan094@gmail.com

ABSTRACT - Protein-based drug target identification, as a novel approach, has found its importance in controlling the threat of emerging antimicrobial resistance, especially in opportunistic microorganisms such as Pseudomonas aeruginosa. Recent advancements in artificial intelligence (AI) and machine learning (ML) have enabled the investigation of large-scale protein data sets, facilitating protein identification and functional annotation. In this article, the authors have proposed a hybrid computational approach, combining unsupervised and supervised machine learning methods, which can be used to analyze the physicochemical properties of P. aeruginosa proteins. Unsupervised methods like K-Means clustering and Principal Component Analysis (PCA) have been integrated into the model to identify the internal patterns among proteins, and Support Vector Machines (SVMs) have been used to classify protein functions. The authors have proven the effectiveness of the AI model in obtaining biologically relevant results regarding protein virulence and resistance with the help of additional protein physicochemical properties, thereby establishing protein analysis as a viable approach in the field of drug discovery.

PAPER INFORMATION

HISTORY
Received: 25 June 2025
Revised: 15 November 2025
Accepted: 25 January 2026
Online: 4 February 2026

MSC
62K05
62K15

KEYWORDS
drug discovery, Machine learning, Target identification, Pseudomonas aeruginosa.

[1]Corresponding Author: [a]Dept. Of CSE, SAGE University, Indore, India, Email: rajdeep_chak@yahoo.co.in

1. INTRODUCTION

The fast development of antimicrobial resistance has become a major worldwide health emergency because it makes current antibiotics less effective, which leads to rising death numbers and disease complications across the globe. Pseudomonas athrivesa poses a major threat because it naturally resists treatment while forming protective biofilms and thrives in hospital settings. Scientists working on antimicrobial research now focus on discovering fresh drug targets from the complete set of proteins that the organism produces [1,8].

The standard drug discovery methods require extended periods of time, while they generate high expenses and produce unsuccessful results. The need for better target identification and prioritization methods has driven researchers to start using computational systems that process data for this purpose. The drug discovery process based on proteins uses their physical and chemical properties together with their structural information to find vital proteins that cause disease and need medical intervention [5].

The field of biological research has undergone a complete transformation, as modern AI systems and ML algorithms now possess the ability to handle complex datasets that contain multiple types of information. Scientists use supervised and unsupervised learning algorithms to predict protein functions and perform target classification and protein clustering, which produces readable answers for drug discovery at its initial stages [6,7]. The methods enable scientists to identify essential patterns in protein data through their own methods instead of depending on experimental data annotations.

The current research presents an AI-based system for protein drug target detection in Pseudomonas aeruginosa bacteria. The research team uses physicochemical protein descriptors together with clustering and classification methods to establish a system that supports functional annotation and target prioritization for antimicrobial drug discovery through computational approaches.

2. LITERATURE REVIEW

Scientists have conducted extensive research on machine learning methods, which they applied to analyze proteins and discover new pharmaceutical drugs. The first research studies proved that ML-based QSAR models together with target prediction systems could identify important biological patterns through their analysis of chemical and protein data [2,6]. Scientists analyze protein structures through unsupervised learning methods, which combine clustering techniques with dimensionality reduction approaches to identify proteins that share similar physical and biological characteristics. The research by Zdrazil and Guha [10] demonstrated that scientists can discover medicinal chemistry and protein–ligand interaction patterns through their scaffold and feature-based analysis approach. Researchers apply PCA and clustering methods to decrease data complexity while they maintain essential biological variation, which helps scientists classify proteins and create visual representations of their data.

Researchers have applied supervised learning methods, which include Support Vector Machines (SVM), to predict protein functions through bioinformatics classification tasks. SVMs function effectively with biological data that contain multiple dimensions because they produce successful results when they predict protein functions through a combination of sequence information and physical chemical properties [4-9].

Scientists have conducted various studies about infectious diseases to find essential proteins and virulent factors that bacteria use to cause disease, so they can develop new medications. Scientists studying Pseudomonas aeruginosa have determined that protein characteristics, including molecular weight and hydrophobicity, and isoelectric point, help researchers understand bacterial resistance and disease-causing abilities [1-3]. The research builds on past investigations by developing a single AI-based system that combines unsupervised protein discovery with supervised P. aeruginosa protein identification. The method uses physicochemical descriptors and scalable ML techniques to create a strong and understandable system that predicts protein-based drug targets.

3. PROPOSED METHODOLOGY

3.1 Proposed Model

This section explains the proposed model as illustrated in **Figure 1**, which outlines the detailed steps followed in the study. The chart presents a hybrid machine learning workflow that combines unsupervised learning techniques—such as clustering and Principal Component Analysis (PCA)for exploratory data analysis, with supervised learning using Support Vector Machine (SVM) for protein function prediction. This pipeline facilitates the classification of proteins and contributes to the identification of potential drug targets and functional annotation in Pseudomonas aeruginosa.

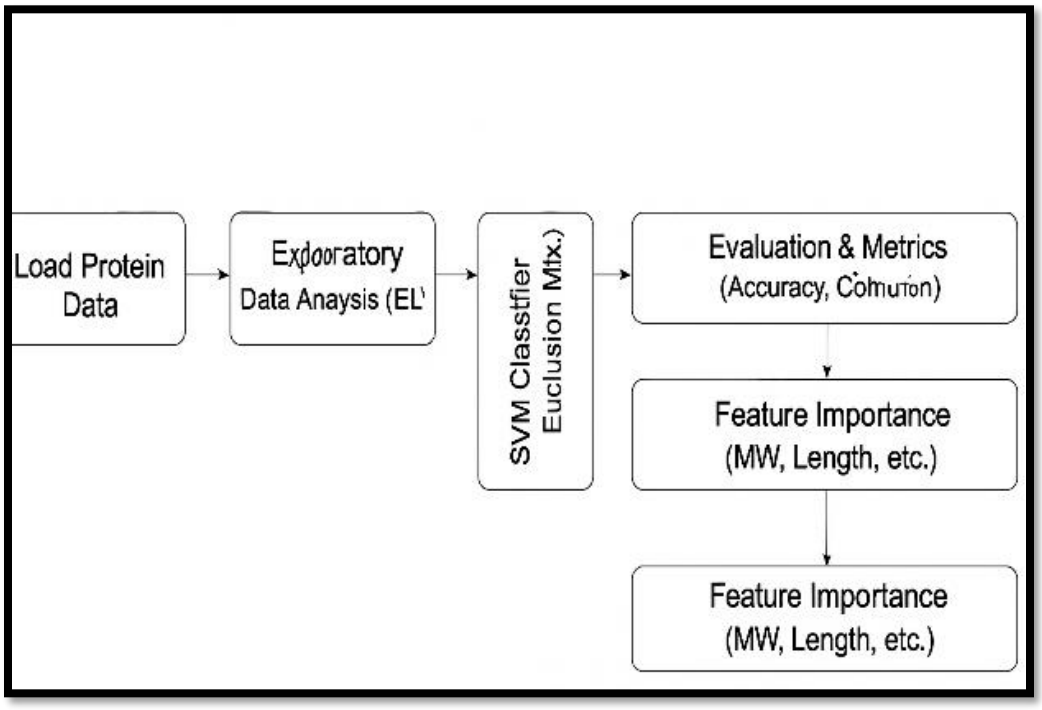

Figure 1. Protein Function Prediction Proposed Model

3.2 Dataset Description

The research study used a dataset that contained physicochemical information about Pseudomonas aeruginosa protein structures [11]. The molecular characteristics of these proteins include their molecular weight and their length, their isoelectric point and their hydrophobic nature, their aromatic content, and their aliphatic index. Scientists use these features as standard bioinformatics tools to study protein characteristics because these features connect to how proteins function and where they exist in cells, and their roles in virulence and antimicrobial resistance.

The dataset contains selected attributes that represent both the structural and functional aspects of proteome diversity, thus creating an ideal dataset for machine learning analysis and drug target identification. The dataset includes information about amino acid sequences together with their molecular weights and isoelectric points, protein lengths, and amino acid compositions and hydrophobic values for various proteins that Pseudomonas aeruginosa produces. The Gram-negative bacterium Pseudomonas aeruginosa causes hospital-acquired infections because it has developed antibiotic resistance, which makes treatment difficult.

This dataset offers several advantages:

The dataset provides new information that helps scientists learn more about bacterial protein characteristics and their operational patterns. Research teams can use this resource to predict protein functions and discover new drug targets that help medical science progress. The system enables researchers to investigate P. aeruginosa antibiotic resistance processes. The dataset serves as a fundamental base that supports different bioinformatics and computational biology applications that focus on protein analysis and drug discovery. The dataset attributes appear in **Table 1** with their corresponding descriptions, which offer complete details about them.

Table 1. Physicochemical Properties of Pseudomonas aeruginosa Proteins

Attribute	Description
ID	A unique identifier assigned to each protein.
Name	The descriptive name or functional annotation of the protein.
Sequence	The amino acid sequence is represented by one-letter amino acid codes.
Molecular _Weight	The molecular weight of the protein is expressed in Daltons.
Isoelectric_Point	The pH value at which the protein carries no net electrical charge.
1. Protein Length	The total number of amino acid residues in the sequence.
Amino_Acid_Composition	A dictionary-style representation of the frequency of each amino acid in the sequence.
Hydrophobicity	A numerical index indicating the hydrophobic character of the protein sequence.

Table 2 shows the statistical characteristics of the dataset attributes. The dataset comprises a total of 1000 proteins with varying molecular weights, sequence lengths, isoelectric points, and hydrophobicity indices. The diversity in physicochemical properties ensures the dataset's suitability for evaluating computational models aimed at predicting protein characteristics in Pseudomonas aeruginosa.

Table 2. Descriptive Statistics of Pseudomonas aeruginosa Protein Dataset

Attribute	Number of Proteins	Mean	Min	Max
Molecular Weight (Da)	1000	40594.47	2461.02	312401.89
Isoelectric Point (pH)	1000	7.14	4.05	11.99
Protein Length (Residues)	1000	364.78	22	2851
Hydrophobicity	1000	-0.35	-1.24	0.91

Figure 2 provides a comprehensive overview of the physicochemical properties of Pseudomonas aeruginosa proteins, allowing researchers to understand the molecular weight, isoelectric points, lengths, and hydrophobicity profiles of these proteins. This information is crucial for further studies in protein function, interactions, and applications in biotechnology.

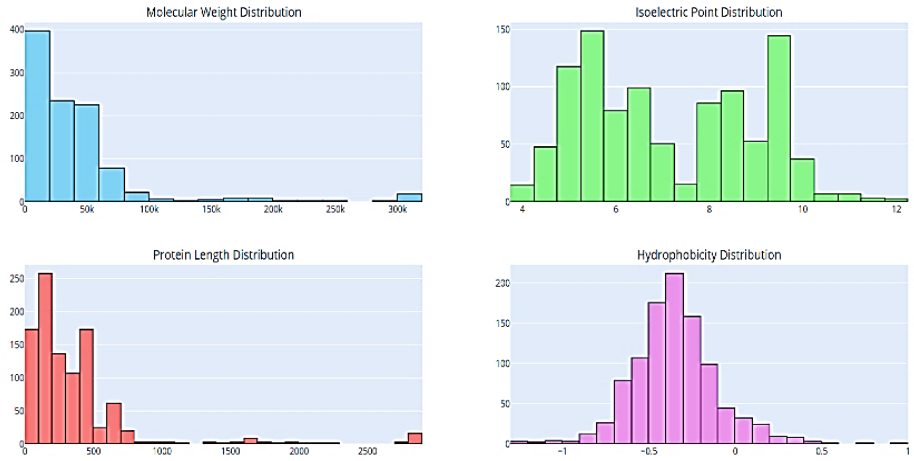

Figure 2. Distribution of Physicochemical Properties of Pseudomonas aeruginosa Proteins

As shown in **Figure 3**, the heatmap provides insights into the interrelationships among these physicochemical properties. The strongest correlation is observed between protein length and molecular weight, while the other correlations are relatively weak. This analysis helps researchers understand the relationships between protein characteristics, which can be critical for protein function and behavior in biological systems.

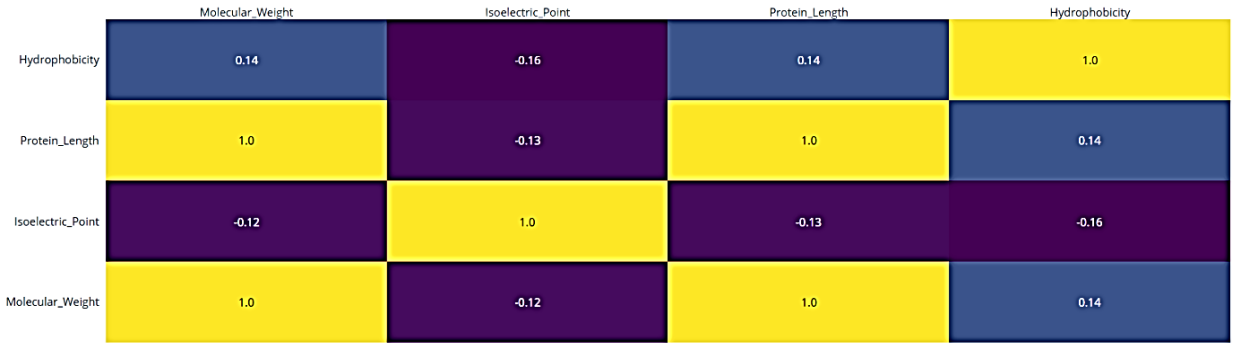

Figure 3. Correlation Heatmap of Physicochemical Properties

3.3 Data Preprocessing

Prior to model development, several **preprocessing steps** were applied to ensure the quality, consistency, and suitability of the protein dataset for machine learning. These steps help clean raw biological data, reduce noise, and enhance the performance of clustering and classification algorithms. The key preprocessing steps are described below:

1. Data Cleaning

The first stage of preprocessing requires users to clean their original protein data sets. The process of handling missing data involves either statistical imputation through mean or median substitution or the complete removal of data points that lack essential molecular weight and hydrophobic information. The process of identifying duplicate protein entries results in their removal from the dataset, which protects against duplicate information and prevents research bias. The dataset receives an evaluation for all its features to verify correct data type assignments because numerical attributes need proper formatting for upcoming machine learning operations [12].

2. Feature Selection

To improve model accuracy and efficiency, only the most relevant features are retained. The research focuses on six essential physicochemical characteristics, which consist of molecular weight, isoelectric point, and hydrophobicity, protein length, aliphatic index, and aromaticity. The process removes all irrelevant and duplicate attributes, which helps the model to reduce noise and prevent overfitting while it uses important biological and computational features for protein classification [12].

3. Feature Scaling / Normalization

The selected features require standardization because their different scales need to be standardized for comparison purposes. The data undergoes Z-score normalization, which transforms every feature through the process of subtracting its average value and then dividing by its standard deviation. The process guarantees that all features, including molecular weight and isoelectric point, will have equal influence during clustering and classification because distance-based algorithms like K-Means and margin-based classifiers like SVM need this type of standardization [12].

4. Outlier Detection and Handling

Biological datasets contain extreme values that produce unpredictable model results. Outlier detection methods work by applying Interquartile Range (IQR) filtering and Z-score threshold systems. The identified outliers, which include extreme protein lengths and molecular weights, are removed or transformed through capping to reduce their influence on clustering and classification processes [12].

5. Dimensionality Reduction (PCA)

The analysis of biological data with many dimensions requires high computational resources because scientists cannot see its complete structure. Principal Component Analysis (PCA) is used to transform original features into fewer features that keep most of the data's original variance. The process becomes faster because of this step, which allows scientists to view protein clusters in two or three dimensions for better biological understanding and improved model development [12].

6. Label Encoding

Users need to transform categorical output classes into numbers when they work with Support Vector Machines (SVM) for supervised learning tasks, which use classes like enzyme, transporter, and hypothetical protein. The algorithm

learns protein functions more effectively through label encoding because it converts protein function classes into unique numbers, which it can process. The algorithm learns protein functions through proper training and produces accurate predictions because it treats class labels as numbers in its system [13], as shown in **Table 3**.

Table 3. Summary of Preprocessing Steps

Step	Description
Data Cleaning	Remove missing, duplicated, or inconsistent entries
Feature Selection	Choose relevant protein features (e.g., MW, pI, hydrophobicity)
Feature Scaling	Normalize data to equalize feature impact
Outlier Handling	Detect and manage anomalies that can skew results
Dimensionality Reduction	PCA is used for visualization and compact representation
Label Encoding	Convert class labels to numeric values for SVM

3.4 Unsupervised Learning: Clustering and Dimensionality Reduction

The research team applied unsupervised learning to discover natural protein dataset patterns because they chose to stay away from using established functional labels. The main objective of the study involved creating groups for Pseudomonas aeruginosa proteins through their physical chemical characteristics which consisted of their molecular weight and their sequence length and their isoelectric point and their hydrophobicity [14].

- **K-Means Clustering**

The K-Means clustering method grouped proteins through their physicochemical properties which scientists used to create similarity-based clusters. The researchers applied the elbow method and silhouette score analysis to find the best cluster count, which produced data groupings that matched the true structure of the dataset. The clusters seem to represent functional categories together with their subcellular locations, which span membrane-bound proteins, cytoplasmic proteins, and extracellular proteins. Scientists can generate useful protein behavior theories from these clusters because they do not need to create any initial labels for their research [15-20].

- **Dimensionality Reduction with PCA**

Scientists used PCA before clustering to help them visualize data while they reduced the complexity of high-dimensional information. The analysis maintained most of the dataset variance, which enabled scientists to visualize protein distributions using two-dimensional or three-dimensional representations. The method improved clustering result understanding because it helped scientists find natural protein clusters that probably share biological functions[16].

- **Applications and Insights**

The unsupervised analysis produces important findings that researchers need for their work in drug discovery and protein research. Scientists can find their next research targets through protein cluster exploration because this method shows how proteins naturally assemble into groups based on their biological roles.

3.5 Supervised Learning: Protein Function Classification

After exploratory clustering, the supervised learning approach was used in classifying proteins and their respective functional groups. For this purpose, the Support Vector Machine (SVM) with RBF Kernel was used to handle high-dimensional biological data and explore the non-linear relationships between physicochemical properties and protein functions.[18]

In this research, the supervised learning approach was used to classify proteins and medical images like DPR, CBCT, and X-rays, which were already labeled. This approach was used to classify new, unseen data based on the patterns learned from the existing data. The supervised learning approach was used on protein data and medical images, which were already labeled. This approach was used to classify unseen data based on existing patterns. The dataset was used to divide it into training and validation sets, and different machine learning and deep learning algorithms like CNN, AlexNet, VGG16, Faster R-CNN, and SVM were used [17].

- **Model Training and Optimization**

The grid search method combined with cross-validation helped me find the best values for the regularization parameter (C) and kernel coefficient (gamma), which brought the model to its highest level of performance. The evaluation process for the trained model involved standard metrics, which included accuracy, together with precision, recall, and F1-score.

- **Overview of Supervised Learning**

The training process of supervised learning models depends on labeled datasets, which contain data points that include their corresponding correct answers. The model learns how to relate input features to their target labels through this process, which enables it to identify unknown data based on its learned patterns.

- **Application in Protein Analysis**

In this study, supervised learning was applied to the labeled protein data set. Each protein is labeled according to its category of function, and this enables the SVM model to learn the mapping between the physicochemical features and the functional outcome. The data set is divided into training and validation sets to enable iterative learning and evaluation of the model for reliable classification of the proteins according to their respective functions.

Table 4 summarizes the stages of the supervised learning pipeline as utilized in the model.

Table 4: Supervised Learning Process in the Proposed Model

Step	Description	Example in the Model
1	Input image	DPR/CBCT scan of a dental structure
2	Label	“Infected”, “Not Infected”
3	Feature Extraction	CNN extracts spatial and structural patterns
4	Classification	SVM or fully connected (FC) layer classifies
5	Evaluation	Accuracy, Precision, Sensitivity, Recall

3.6 Hyperparameters Tuning

In the proposed model, several hyperparameters are used to optimize performance for both SVM and K-Means algorithms. For the Support Vector Machine (SVM), key hyperparameters include C, which controls the trade-off between model complexity and training accuracy; kernel, which defines the type of decision boundary (e.g., linear, RBF, or polynomial); gamma, which determines the influence of individual data points in non-linear kernels; and degree, which applies when using a polynomial kernel. In the K-Means clustering component, important hyperparameters are n_clusters, specifying the number of clusters to form; init, which determines the method for initializing centroids (such as 'k-means++'); max_iter, setting the maximum number of iterations allowed; n_init, defining how many times the algorithm will run with different initializations; and random_state, used to ensure reproducibility. These hyperparameters are carefully selected and tuned to ensure accurate protein classification and meaningful clustering results [19].

3.7 Network architecture

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification tasks, particularly effective in high-dimensional and complex biological datasets. In the context of this study, SVM serves as the core classification technique for predicting the functional class of *Pseudomonas aeruginosa* proteins based on their physicochemical properties. These properties include molecular weight, protein length, isoelectric point, and hydrophobicity, among others. As shown in **Figure 4**, SVM operates by identifying an optimal hyperplane that separates data points of different classes with the maximum possible margin. When the data is not linearly separable

in its original space, kernel functions, such as radial basis function (RBF) or polynomial kernels can be employed to project the data into a higher-dimensional space where a linear separation is feasible. This makes SVM particularly suited for handling non-linear and overlapping biological data.

In the proposed model, the SVM classifier is trained using labeled protein data to learn the relationship between input features and protein function. Once trained, the model can predict the class of unseen proteins, enabling functional annotation based solely on physical and chemical characteristics. The model demonstrates strong performance in frequently occurring classes and highlights key predictive features, notably molecular weight and protein length. However, classification accuracy tends to decline for rare or underrepresented protein classes, reflecting a common challenge in imbalanced biological datasets. Overall, the integration of SVM into the model provides a robust, interpretable, and efficient method for classifying protein function, contributing to the broader goal of facilitating protein-based drug discovery in Pseudomonas aeruginosa.

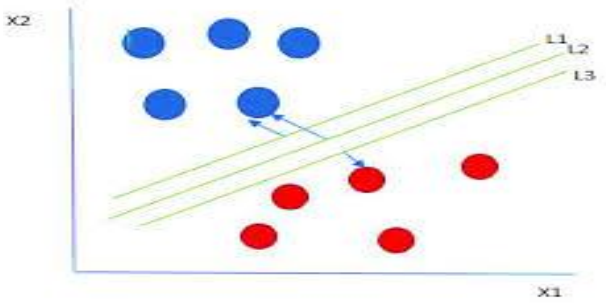

Figure 4. Multiple hyperplanes separate the data from two classes

4. RESULTS AND DISCUSSION

4.1 Exploratory Analysis and Clustering Results

The dataset represents Pseudomonas aeruginosa proteins and their important physicochemical properties, such as molecular weight, isoelectric point, hydrophobicity, aromaticity, aliphatic index, and protein length **Table 4**. A statistical description of these properties is given in **Table 1**, ensuring that there is enough variability for machine learning analysis.

Before analysis, the dataset was preprocessed for missing values, Z-normalization, and removal of outliers to ensure consistency and validity. Principal Component Analysis (PCA) was performed to reduce dimensionality while retaining variance, where the first two to three components explained more than 80% of the total variance, allowing for effective visualization. The PCA plots showed natural protein clusters based on the physicochemical diversity of the P. aeruginosa proteome.

Later, K-Means clustering was employed to explore natural structures in the data. The number of clusters was determined by the elbow method and silhouette analysis. The clusters were formed based on proteins with similar biochemical properties, such as molecular weight, protein length, and hydrophobicity, with some clusters indicating possible membrane-bound or virulence-related proteins **Figure 5.**

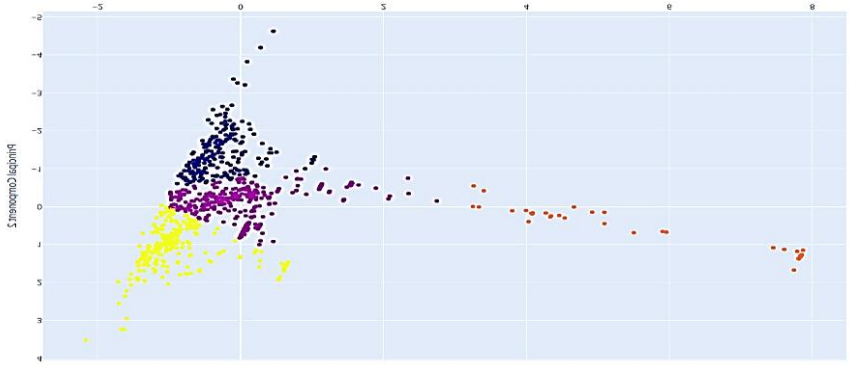

Figure 5. K-Means clustering of Proteins

The performance of the SVM classifier in terms of the prediction of protein function shows that the model performs well in the prediction of most of the protein function classes. High accuracy and F1 scores were obtained for the protein function classes that were well represented in the dataset, showing that the physicochemical features offer sufficient discriminative ability for protein function prediction. Misclassification of protein function is mainly due to the presence of similar physicochemical features, which is a common limitation in the classification of biological data. However, the performance of the model shows that SVM is a good classifier for protein-based classification in drug discovery research.

The classification phase, as shown in **Figure 6**, involved training a Support Vector Machine (SVM) with a radial basis function (RBF) kernel on labeled data. The model was optimized using hyperparameters such as C, gamma, and kernel, and evaluated using metrics like accuracy, precision, recall, and F1-score. The SVM achieved high classification performance, correctly predicting the functional classes of most proteins. The confusion matrix confirmed that misclassifications were minimal and typically occurred between proteins with similar structural properties.

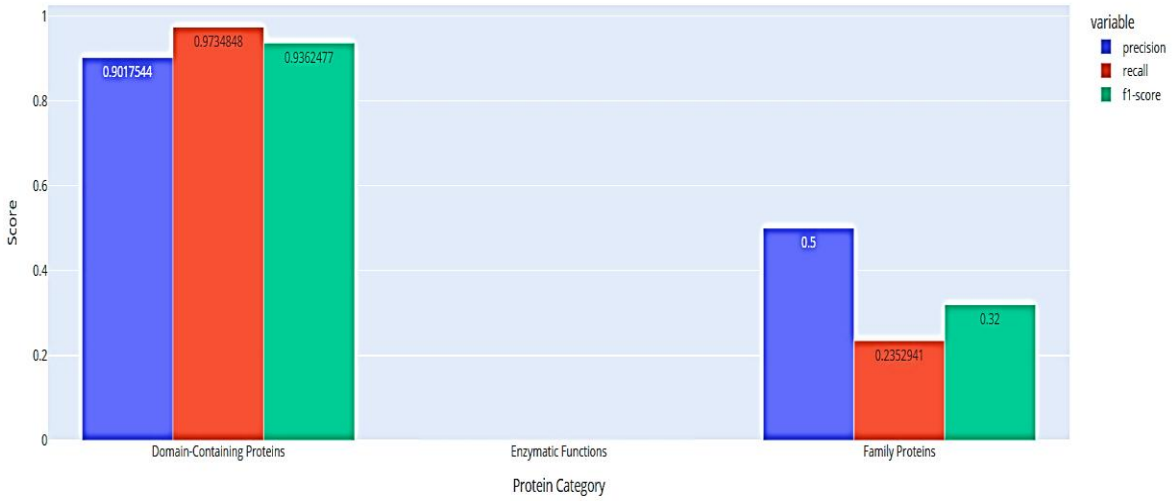

Figure 6. Classification Report Metrics by Class

Lastly, the results were assessed for biological relevance. Proteins grouped in specific clusters or classified into functional classes showed consistency with known annotations. Many of the predicted virulence factors and essential proteins exhibited traits associated with drug targets or antibiotic resistance mechanisms, such as efflux pumps or biofilm-forming proteins. This demonstrates the model's potential not only in prediction but also in generating biological hypotheses for future drug development efforts.

5. Conclusion and Future Directions

The current study offers a comprehensive AI-based framework for protein-based drug target predictions for the pathogen Pseudomonas aeruginosa, addressing the critical problem of drug target predictions for antimicrobial drugs. By using physicochemical protein descriptors along with unsupervised and supervised machine learning methods, the current study offers an efficient method for the exploration, classification, and prioritization of proteins of potential therapeutic value.

The integration of K-Means clustering and PCA analysis enabled the identification of the intrinsic groups of proteins, revealing patterns of biological relevance concerning the structure, localization, and function of proteins. These patterns were further reinforced by the implementation of Support Vector Machine-based classification, which exhibited significant efficiency in the prediction of the functional classes of proteins based on physicochemical features. Furthermore, the current study found that certain groups of proteins were associated with virulence factors and resistance mechanisms, thus proving their potential as drug targets. The current study offers significant methodological value concerning the application of machine learning methods for the exploration of proteins of potential therapeutic value. Unlike the conventional methods of drug target predictions, the current study offers a cost-effective and reproducible method of drug target predictions for the pathogen Pseudomonas aeruginosa, which is associated with complex drug resistance.

It is also possible for future research to build upon this approach by incorporating more protein descriptors, features, and protein-protein interactions, which can improve the accuracy of the predictions. Additionally, it is also possible for this proposed pipeline to be combined with other sophisticated deep learning architectures and other biological knowledge sources to aid in the dynamic prioritization of the targets and the discovery of new antimicrobial compounds. Overall,

this research shows the potential of AI-based protein analysis as a powerful tool for aiding the fight against antimicrobial resistance.

ACKNOWLEDGMENTS

The authors sincerely thank the referees, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which have greatly improved this paper. The authors also acknowledge the use of DeepSeek for assistance in improving the English grammar and language clarity.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCE

[1] Antolin, A.A., Workman, P., & Al-Lazikani, B. (2021). Public resources for chemical probes: The journey so far and the road ahead. Future Medicinal Chemistry, 13(8), 731–747. https://doi.org/10.4155/fmc-2020-0306
[2] Bosc, N., Atkinson, F., Félix, E., Gaulton, A., Hersey, A., & Leach, A.R. (2019). Large-scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. Journal of Cheminformatics, 11, 4. https://doi.org/10.1186/s13321-018-0325-4
[3] Breidenstein, E.B.M., de la Fuente-Núñez, C., & Hancock, R.E.W. (2011). Pseudomonas aeruginosa: All roads lead to resistance. Trends in Microbiology, 19(8), 419–426. https://doi.org/10.1016/j.tim.2011.04.005
[4] Hanser, T., Steinmetz, F.P., Plante, J., et al. (2019). Avoiding hERG liability in drug design via QSAR and data fusion. Journal of Cheminformatics, 11, 9. https://doi.org/10.1186/s13321-019-0338-7
[5] Leeson, P.D., Bento, A.P., Gaulton, A., et al. (2021). Target-based evaluation of drug-like properties and ligand efficiencies. Journal of Medicinal Chemistry, 64(11), 7210–7230. https://doi.org/10.1021/acs.jmedchem.0c02117
[6] Mayr, A., Klambauer, G., Unterthiner, T., et al. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chemical Science, 9, 5441–5451. https://doi.org/10.1039/C8SC00148K
[7] Merk, D., Friedrich, L., Grisoni, F., & Schneider, G. (2018). De novo design of bioactive small molecules by artificial intelligence. Molecular Informatics, 37(1–2), 1700153. https://doi.org/10.1002/minf.201700153
[8] Oprea, T.I., Bologa, C.G., Brunak, S., et al. (2018). Unexplored therapeutic opportunities in the human genome. Nature Reviews Drug Discovery, 17(5), 317–332. https://doi.org/10.1038/nrd.2018.14
[9] Walter, M., Allen, L.N., de la Vega de León, A., et al. (2022). Benefits of imputation models over traditional QSAR for toxicity prediction. Journal of Cheminformatics, 14, 32. https://doi.org/10.1186/s13321-022-00615-3.
[10] Zdrazil, B., & Guha, R. (2018). The rise and fall of a scaffold: A trend analysis of scaffolds in the medicinal chemistry literature. Journal of Medicinal Chemistry, 61(11), 4688–4703. https://doi.org/10.1021/acs.jmedchem.7b01631.
[11] UniProt Consortium. (2018). UniProt proteome database entry UP000253594. Submitted to EMBL/GenBank/DDBJ databases (July 2018). Retrieved 2026, from https://www.uniprot.org/proteomes/UP000253594.
[12] Koukaras, P., & Tjortjis, C. (2025). Data preprocessing and feature engineering for data mining: Techniques, tools, and best practices. AI, 6(10), 257. https://doi.org/10.3390/ai6100257.
[13] Kaliappan, J., Saravana Kumar, I.J., et al. (2024). Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. Frontiers in Artificial Intelligence, 7, 1421751. https://doi.org/10.3389/frai.2024.1421751.
[14] Kuo, C.-T., Xu, D., & Friesen, R. (2025). A brief review of unsupervised machine learning algorithms: Dimensionality reduction and clustering. Universe, 11(12), 412. https://doi.org/10.3390/universe11120412.
[15] Kamdar, N., & Musen, M.A. (2025). Exploring homology detection via k-means clustering of proteins embedded with a large language model. Bioinformatics, 41(10). https://doi.org/10.1093/bioinformatics/btaf472.
[16] Kılıç, D.K., & Nielsen, P. (2023). Comparative analyses of unsupervised PCA–KMeans algorithm. Sensors, 22(23), 9172. https://doi.org/10.3390/s22239172.
[17] deHealth Lab. (2024). An overview on the advancements of support vector machine models in healthcare applications: A review. Information, 15(4), 235. https://doi.org/10.3390/info15040235.
[18] Zhang, Y., et al. (2024). A review of machine learning techniques for the classification and detection of breast cancer from medical images. Diagnostics, 13(14), 2460. https://doi.org/10.3390/diagnostics13142460.
[19] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[20] Montgomery, R.M. (2024). Overview of clustering techniques: From k-means to spectral methods.