

# **Functional Protein Prediction Using Machine Learning**

Arghyadeep Das

B.Tech. Biotechnology, 2<sup>nd</sup> Year

Adamas University

Mentor Name: Adrija Das

Period of Internship: 14th Jan 2025 - 30th April 2025

Report submitted to: IDEAS – Institute of Data  
Engineering, Analytics and Science Foundation, ISI  
Kolkata

## Abstract

This study presents a machine learning-based approach for predicting protein function utilizing physicochemical features derived from a curated dataset of *Pseudomonas aeruginosa* proteins. Physicochemical parameters, including molecular weight, hydrophobicity, isoelectric point, and instability index, were employed as predictive features. Rigorous data cleaning was performed to eliminate biologically irrelevant or sparsely represented data, ensuring model robustness.

Among the evaluated models, a Random Forest classifier implemented within a One-vs-Rest (OVR) framework demonstrated superior classification accuracy (91.6%), outperforming XGBoost (87%) and other baseline models (Fig 9). Functional interpretation of the predictions was achieved through a multi-faceted annotation strategy employing Gene Ontology (GO) term mapping via GOATOOLS, protein-protein interaction (PPI) network analysis using STRING-DB, and pathway mapping with KEGG. The integration of these resources provided a biologically consistent framework for understanding the functional implications of the machine learning predictions, highlighting the potential of physicochemical-based approaches for protein function annotation, particularly in understudied organisms.

## Introduction

The exponential increase in sequenced genomes has shifted the primary challenge in bioinformatics from sequence acquisition to functional annotation. While protein sequences are readily available, elucidating their biological roles remains a critical hurdle. Traditional sequence homology-based function prediction methods often prove inadequate for novel proteins or those lacking close homologs with known functions. An alternative strategy leverages the inherent relationship between a protein's physicochemical properties and its functional attributes, as these characteristics often correlate with structural and functional roles.

This study investigates the utility of supervised machine learning models trained on the physicochemical properties of *Pseudomonas aeruginosa* proteins for predicting their functional classes, independent of primary sequence information. This approach holds particular significance in contexts where sequence data is abundant but functional annotations are limited, offering a means to generate testable hypotheses for subsequent experimental validation.

## Project Objectives

- To predict functional classes of proteins using physicochemical parameters.
- To apply and evaluate machine learning models for biological classification.
- To map predicted classes to GO terms using GOATOOLS.
- To contextualize predictions through STRING-DB PPI networks and KEGG metabolic pathways.

## Methodology

### Data Source

The dataset utilized in this study was obtained from Kaggle and it is mentioned in the appendix. It encompasses physicochemical and structural properties of proteins derived from *Pseudomonas aeruginosa*. The dataset (Fig 1) consists of 1,000 protein entries, each characterized by the following

- ID: A unique identifier for each protein entry.
- Name: The name of the protein.
- Sequence: The amino acid sequence of the protein
- Molecular Weight: The calculated molecular weight of the protein in Daltons
- Isoelectric Point (pI): The pH at which the protein carries no net electric charge
- Protein Length: The number of amino acids in the protein sequence,
- Amino Acid Composition: The frequency or percentage of each amino acid within the protein sequence.
- Hydrophobicity: A measure of the protein's hydrophobic (water-repelling) properties

ID	Name	Sequence	Molecular_Weight	Isoelectric_Point	Protein_Length	Amino_Acid_Composition	Hydrophobicity
WP_369686368.1	ATP-binding cassette domain-containing protein, part	MLELNFTQLGSH	5756.543	8.517643547	56	{'M': 1, 'L': 8, 'E': 2, 'N': 3, 'F': 2,	0.339285714
WP_369686367.1	aldehyde dehydrogenase family protein, partial [Pseud	cMQSRDNGKPLAE	6617.5065	6.106917763	62	{'M': 2, 'Q': 2, 'S': 3, 'R': 6, 'D': 3,	-0.146774194
WP_369686366.1	hypothetical protein, partial [Pseudomonas aeruginosa	GGEYLEIIEAARDIF	9303.2892	4.533444023	81	{'G': 4, 'E': 8, 'Y': 3, 'L': 9, 'I': 6, 'V	-0.40617284
WP_369686365.1	hypothetical protein, partial [Pseudomonas aeruginosa	NAVVNQKRVLPAF	6304.0708	9.989714622	58	{'N': 5, 'A': 5, 'V': 4, 'Q': 3, 'K': 5,	-0.59137931
WP_369686364.1	homocysteine S-methyltransferase family protein [Pse	MAGYLPQWLDAG	3619.1997	7.810425377	34	{'M': 1, 'A': 5, 'G': 4, 'Y': 1, 'L': 4,	0.141176471
WP_369686363.1	hypothetical protein, partial [Pseudomonas aeruginosa	NRLILSPMGVRDV	9303.6417	8.358213234	83	{'N': 4, 'R': 6, 'L': 7, 'I': 7, 'S': 11,	-0.059036145

Fig 1: - Initial dataset having Dimensions 1000 X 6 elements

## Data Cleaning and Preprocessing

Initial data preprocessing involved the removal of non-informative or redundant features, specifically excluding primary sequence data, NCBI ID and amino acid composition.

To ensure robust model training and generalization, protein functional classes (Name column) with fewer than 25 occurrences were excluded from the analysis. We were left with 6 protein classes (Fig 2) in our dataset.

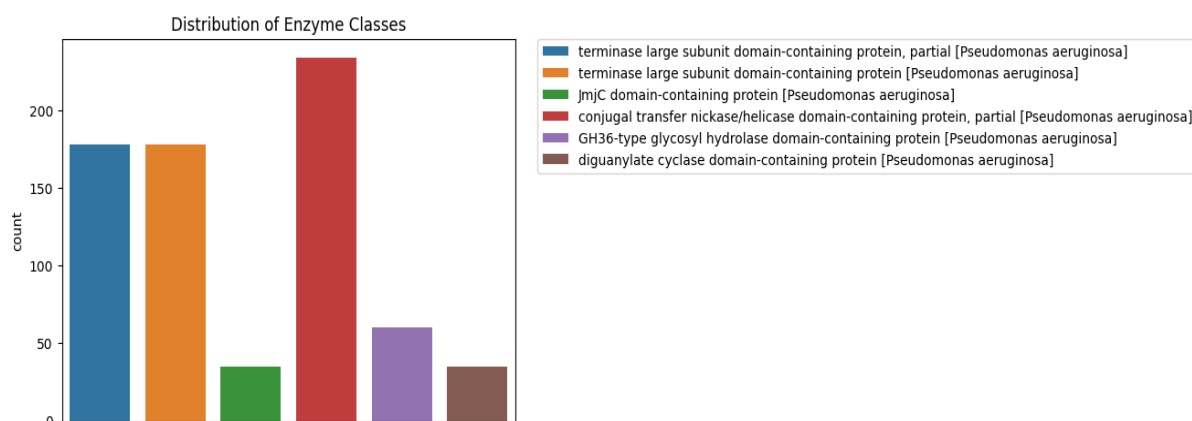


Fig 2: - Protein Class Distribution

The target variable was label-encoded to facilitate machine learning model implementation, and all physicochemical features were converted to numerical data types for computational compatibility.

Labelled Name	Original Name
0	GH36-type glycosyl hydrolase domain-containing protein
1	JmjC domain-containing protein
2	conjugal transfer nickase/helicase domain-containing protein, partial
3	diguanylate cyclase domain-containing protein
4	terminase large subunit domain-containing protein
5	terminase large subunit domain-containing protein, partial

Fig 3: - Original enzyme names mapped to their labelled numbers

Consistent with the principle of maintaining biological realism, no synthetic data augmentation techniques (like SMOTE) were employed. The artificial manipulation of biologically derived datasets can introduce spurious patterns and compromise the biological interpretability of the results.

Name ▾	Molecular_Weight ▾	Isoelectric_Point ▾	Protein_Length ▾	Hydrophobicity ▾
5	33681.6776	9.363275337	291	-0.712371134
5	7948.8123	5.570755577	69	-0.52173913
5	34903.0957	5.249615288	309	-0.287378641
4	48147.0978	8.396185112	428	-0.391588785
5	41588.8357	8.256933403	363	-0.308539945
5	58610.8042	7.612582588	510	-0.489803922
5	15247.097	4.474729347	133	-0.215789474
5	54249.8815	7.120243263	470	-0.518723404

Fig 4:– Final Dataset having Dimension 720 X 5

## EDA and Visualization

To understand the underlying patterns and relationships within the dataset, exploratory data analysis (EDA) techniques were employed

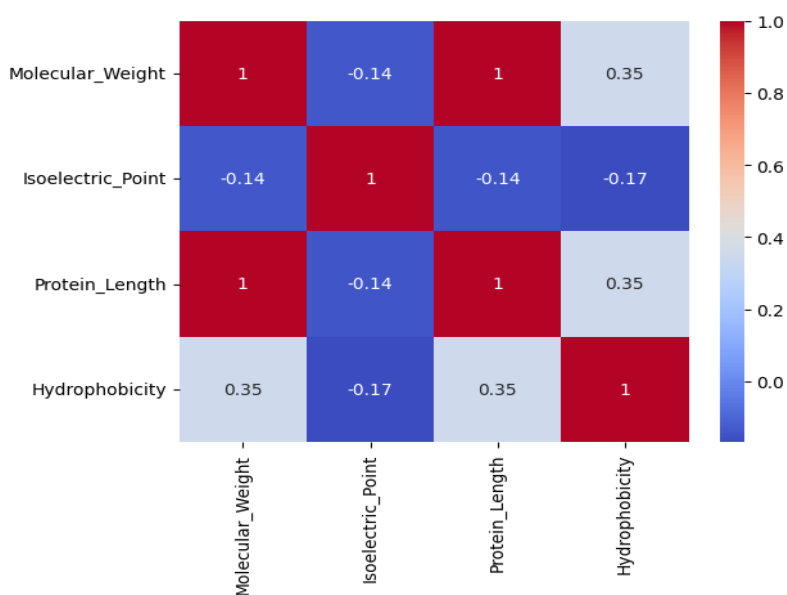


Fig 5: - Correlation Heatmap

A correlation heatmap (Fig 5) was generated to visualize the relationships between numerical features. The colour intensity represents the strength of the correlation, with red indicating positive correlations and blue indicating negative correlations. The correlation coefficients are also displayed within each cell.

Overall, except for the perfect correlation between Molecular Weight and Protein Length, there are no strong linear dependencies between the other pairs of features. This suggests that each of these other features likely contributes some unique information to a model.

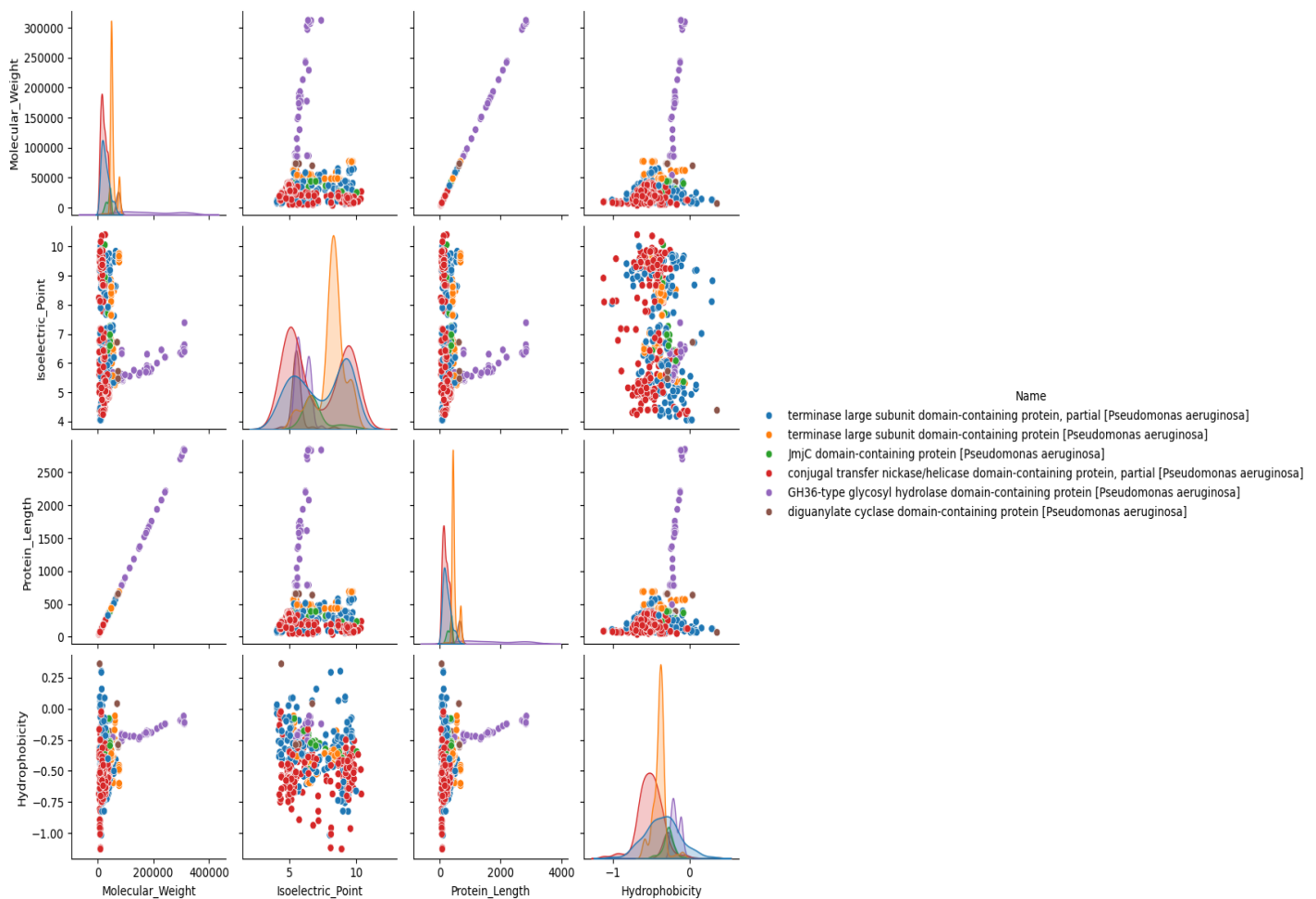


Fig 6: - Pair-plot of features

This pair plot (Fig 6) provides a visual exploration of the relationships between the four physicochemical features (Molecular Weight, Isoelectric Point, Protein Length, and Hydrophobicity) and how they might vary across different protein functional classes

1. **Molecular Weight:** Shows a distribution skewed towards lower values with a few peaks, suggesting that most proteins in the dataset have relatively lower molecular weights, but there are also subgroups with significantly higher weights.
2. **Isoelectric Point:** Proteins in the dataset tend to cluster around several different isoelectric points (around pH 4-6, 6-8, and slightly higher). Different functional classes appear to have varying preferences for these pI ranges.
3. **Protein Length:** As expected, the distribution is skewed towards shorter lengths, with a long tail indicating the presence of some very long proteins. Like molecular weight, different classes show overlaps but might have characteristic length distributions.
4. **Hydrophobicity:** Shows a distribution centered around slightly negative values, suggesting that most proteins in the dataset have a net hydrophilic character. However, there's a significant spread, indicating the presence of proteins with varying degrees of hydrophobicity. Different functional classes seem to exhibit distinct hydrophobicity profiles.

## Feature Engineering

This horizontal bar chart (Fig 7) visualizes the **feature importance scores** as determined by Random Forest model for predicting protein functional classes. Each bar represents a different physicochemical feature, and the length of the bar corresponds to its importance score. The features are listed on the y-axis, and the feature importance score is on the x-axis.

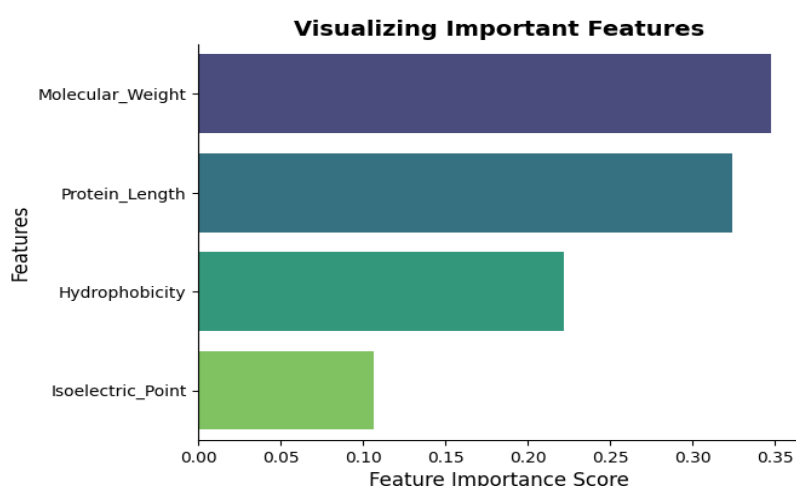


Fig 7: - Feature Importance Graph

The feature importance scores from the trained Random Forest model, indicates the relative contribution of each physicochemical feature in predicting protein functional classes. Molecular Weight and Protein Length emerge as the most important features, followed by Hydrophobicity and then Isoelectric Point. This suggests that the model heavily relies on the size-related properties and the overall hydrophobic character of the proteins to discriminate between different functional categories.

## Model Selection and Training

A suite of classification algorithms, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost, were implemented and evaluated based on their classification accuracy and interpretability.

The Random Forest classifier, implemented within a One-vs-Rest (OVR) strategy to handle the multi-class nature of the target variable, was selected as the optimal model due to its robust performance on high-dimensional datasets and its capacity to model complex non-linear relationships. The dataset was partitioned into training (80%) and testing (20%) sets. Model performance was assessed using standard classification metrics. Following prediction, inverse transformation of the encoded labels was performed to facilitate biological interpretation. The code is provided in the appendix.

## Functional Annotation

To derive biological insights from the predicted functional classes, a multi-stage annotation pipeline was implemented, leveraging existing Gene Ontology (GO) annotations and pathway databases.

### 1) GO Term Annotation Using GOATOOLS

To biologically contextualize the protein class predictions generated by the machine learning model, we employed **Gene Ontology (GO) term annotation** as a structured and standardized approach to describe protein function.

#### **GO Terms**

The **Gene Ontology (GO)** is an internationally recognized bioinformatics initiative that provides a controlled vocabulary to describe gene and protein attributes across species. GO terms are organized under three primary domains:

- **Molecular Function (MF):** Describes elemental activities of a gene product at the molecular level (e.g.-ATP binding, DNA helicase activity)
- **Biological Process (BP):** Refers to biological goals or pathways that a gene product contributes to (e.g., carbohydrate metabolism, glycolysis)
- **Cellular Component (CC):** Indicates where in the cell the gene product is active (e.g., cytoplasm, membrane)

Each GO term is represented by a unique **GO identifier (GO ID)** (e.g.- GO:0030246), a descriptive name, and an associated namespace (MF, BP, or CC).

#### **Purpose in This Study**

While the machine learning classifier successfully predicted functional classes based on physicochemical properties, those class labels alone do not convey **specific biological meaning**. Mapping predicted enzyme classes to GO terms helps to:

- Assign standard **functional descriptors** to model outputs
- Support downstream validation through **biological databases (e.g., STRING, KEGG)**

Thus, GO term annotation bridges the gap between computational classification and biological function.



## Annotation Workflow

We implemented the following procedure to map predicted protein classes to GO terms:

### Inverse Transformation of Predictions:

- The numeric labels output by the model were mapped back to their original enzyme class names (e.g., “GH36-type glycosyl hydrolase domain-containing protein”).

### GO Term Retrieval Using GOATOOLS:

- We used the **GOATOOLS** Python library to query GO annotations manually associated with each enzyme name.
- GOATOOLS interfaces with structured databases (e.g., Gene Ontology Annotation [GOA], UniProt) to retrieve:
  - GO Term Name
  - GO Namespace (MF, BP, CC)

### Data Structuring:

- The retrieved information was compiled into a tabular format:
- **Enzyme Name | GO ID | Function | Namespace**
- This table provided a one-to-one mapping between predicted classes and their inferred biological roles.(Fig 8)

	Enzyme name	GO ID	Function	Namespace
1	GH36-type glycosyl hydrolase domain-contain	GO:0030246	carbohydrate binding	molecular_function
2	GH36-type glycosyl hydrolase domain-contain	GO:0005975	carbohydrate metabolic process	biological_process
3	GH36-type glycosyl hydrolase domain-contain	GO:0004339	glucan 1,4-alpha-glucosidase activity	molecular_function
4	GH36-type glycosyl hydrolase domain-contain	GO:0016757	glycosyltransferase activity	molecular_function
5	JmjC domain-containing protein [Pseudomon	GO:0016706	2-oxoglutarate-dependent dioxygenase activ	molecular_function
6	conjugal transfer nickase/helicase domain-co	GO:0005524	ATP binding	molecular_function
7	conjugal transfer nickase/helicase domain-co	GO:0003677	DNA binding	molecular_function
8	conjugal transfer nickase/helicase domain-co	GO:0003678	DNA helicase activity	molecular_function
9	diguanylate cyclase domain-containing prote	GO:0052621	diguanylate cyclase activity	molecular_function
10	diguanylate cyclase domain-containing prote	GO:0005525	GTP binding	molecular_function
11	diguanylate cyclase domain-containing prote	GO:0046872	metal ion binding	molecular_function
12	diguanylate cyclase domain-containing prote	GO:0007165	signal transduction	biological_process
13	terminase large subunit domain-containing pi	GO:0005524	ATP binding	molecular_function
14				

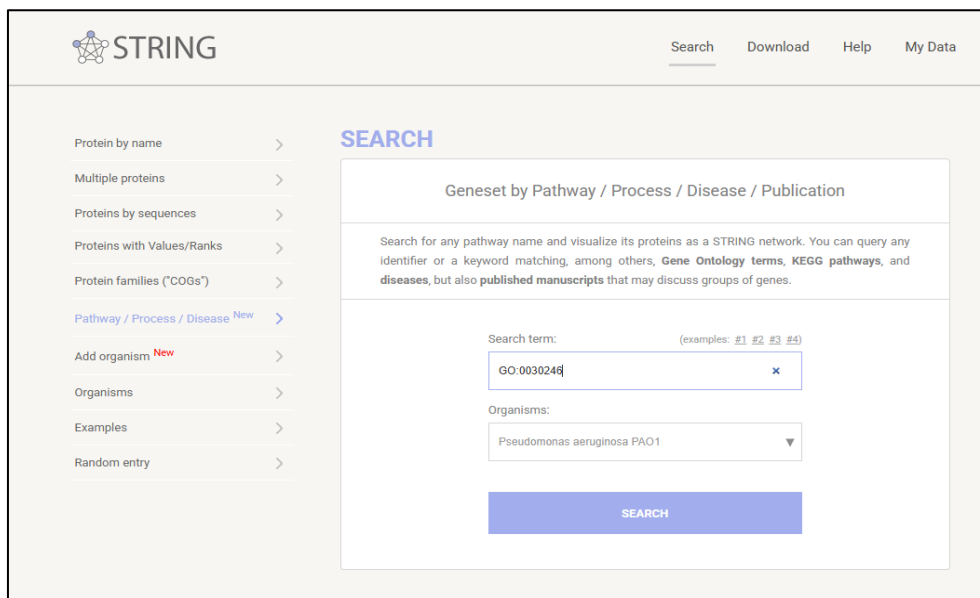
Fig 8: - Name to Go Id's mapped

These GO terms provide immediate biological context to the predictions made by the model.

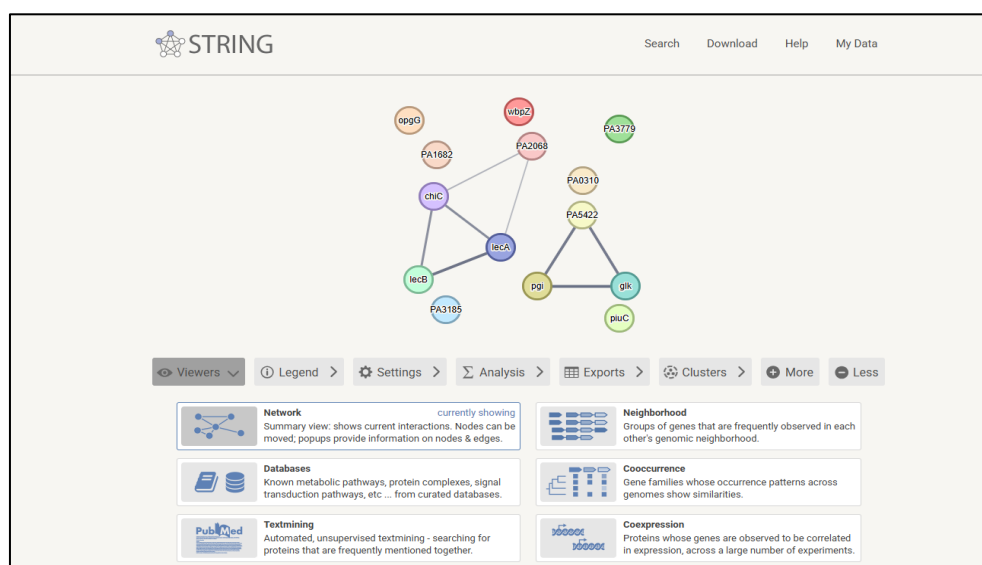
## 2) Functional Network Analysis with STRING-DB

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a widely used database and web resource that provides known and predicted protein-protein interactions (PPIs). The website homepage is mentioned in the appendix.

To understand the potential functional interdependencies between proteins within *Pseudomonas aeruginosa* that share the same GO IDs identified through the goatools analysis, the STRING database was employed. For each selected GO ID, a list of corresponding *Pseudomonas aeruginosa* proteins was generated and subsequently used to construct protein-protein interaction (PPI) networks. Following shows the methods followed



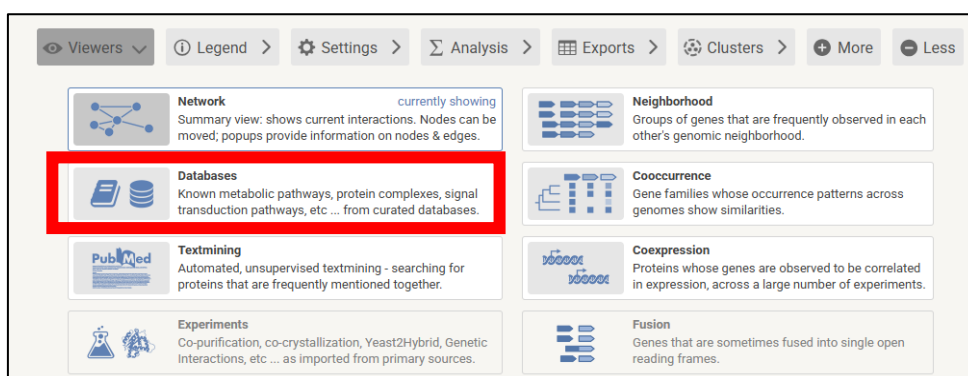
The screenshot shows the STRING database search page. On the left is a sidebar with navigation options: Protein by name, Multiple proteins, Proteins by sequences, Proteins with Values/Ranks, Protein families ("COGs"), Pathway / Process / Disease (highlighted), Add organism (marked as New), Organisms, Examples, and Random entry. The main area is titled "SEARCH" and "Geneset by Pathway / Process / Disease / Publication". It contains a search form with a "Search term:" field (with examples #1, #2, #3, #4) containing "GO:0030246", an "Organisms:" dropdown menu set to "Pseudomonas aeruginosa PAO1", and a "SEARCH" button.



### 3) Pathway Mapping via KEGG

The **Kyoto Encyclopaedia of Genes and Genomes (KEGG)** is a widely used bioinformatics resource that provides curated knowledge on biological systems, including Metabolic pathways. The website homepage is mentioned in the appendix.

To contextualize the functional roles of these *Pseudomonas aeruginosa* proteins at a pathway level, the protein lists generated from the STRING-DB analysis (grouped by shared GO IDs) were mapped onto pathways within the KEGG. This allowed for the visualization of the specific location and potential roles of these proteins within known metabolic and signalling pathways within the organism. Following shows the methods followed



DATABASE KNOWLEDGE	
Relevant datasets in <i>Pseudomonas aeruginosa</i> :	
annotated pathway (KEGG) (kegg pathways) Name: Glycolysis / Gluconeogenesis	pgi glk PA5422 [... and 34 other proteins]
annotated pathway (KEGG) (kegg pathways) Name: Microbial metabolism in diverse environments	pgi glk PA5422 [... and 276 other proteins]
annotated pathway (KEGG) (kegg pathways) Name: Biosynthesis of secondary metabolites	pgi glk PA5422 [... and 391 other proteins]
annotated pathway (KEGG) (kegg pathways) Name: Metabolic pathways	pgi glk PA5422 [... and 969 other proteins]

**Single Interaction Record**

annotated pathway (KEGG) [link out: KEGG](#)

Name: Glycolysis / Gluconeogenesis

*Comment:* Manually curated metabolic and signalling pathways were imported from KEGG (August 2022). Pathway neighbors and subunits of the same enzyme/complex were given association links

***Pseudomonas aeruginosa*:** pgi glk PA5422 pgk fda acsA PA1027 pykF lpdG ppsA exaA exaC lpdV PA2275 PA2323 PA2555 PA3001 gapA PA3415 PA3416 pdhA adhC eno PA4022 PA4152 pykA acsB tpiA lpd3 PA4899 aceE aceF fbp pgm pckA algC adhA

[BACK](#)

**KEGG Glycolysis / Gluconeogenesis - *Pseudomonas aeruginosa* PA01**

[ [Pathway menu](#) | [Organism group](#) | [Pathway entry](#) | [Show description](#) | [Download](#) | [Help](#) ]

[Change pathway type](#)

# Results

## Model Performance

The Random Forest classifier demonstrated a robust ability to predict protein functional classes based solely on physicochemical features, achieving a high accuracy of **91.6%**. (Fig 9). This significantly outperformed other tested algorithms, including XGBoost (87.0%)(Fig 9), suggesting that the ensemble learning approach of Random Forest is particularly well-suited for capturing the complex relationships between these intrinsic protein properties and their functional roles. The lower performance of linear models like Logistic Regression and the more linear boundary-seeking SVM highlights the potentially non-linear nature of the relationship between the chosen physicochemical features and protein functional classification. This strong predictive power underscores the potential of using easily computable protein parameters for initial functional assessments, especially in cases where sequence-based homology searches might be limited.

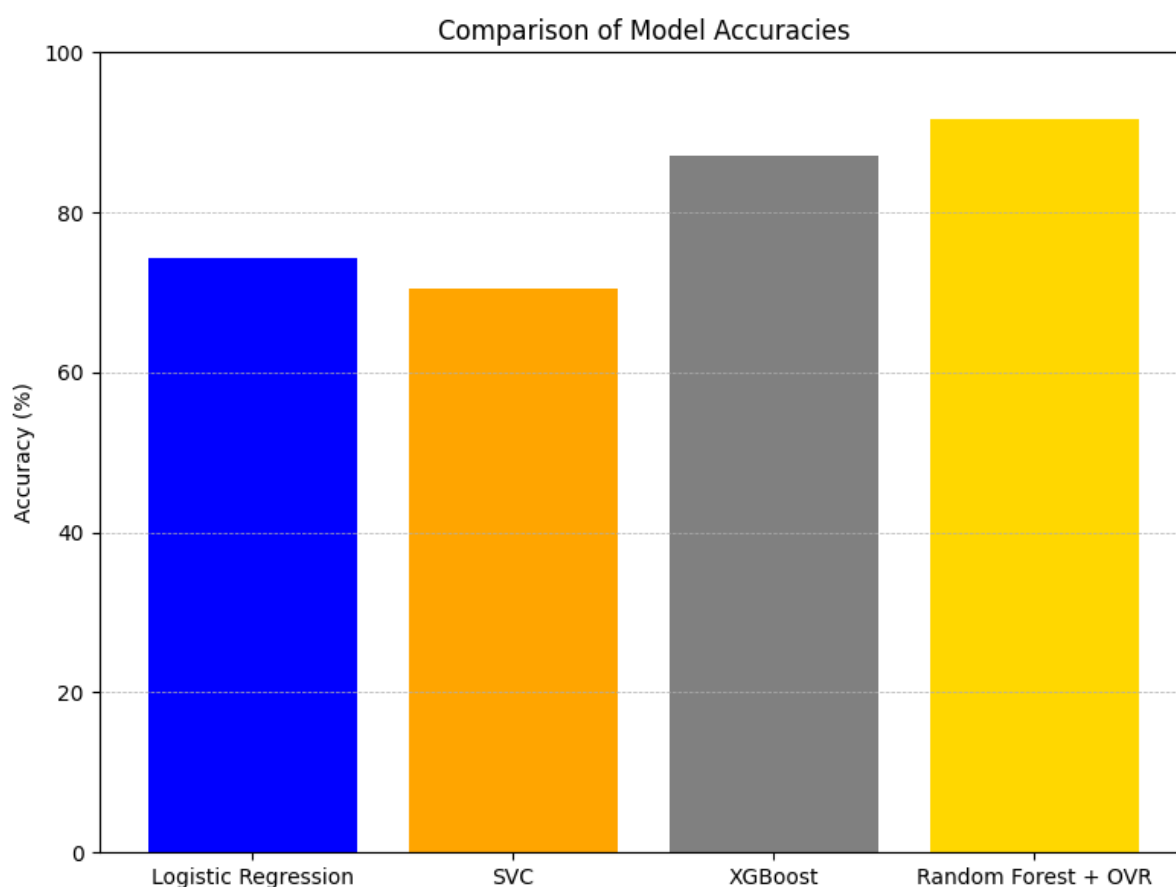


Fig 9: – Accuracy Comparison of Different Models

## Gene Ontology (GO) Analysis

The distribution of GO terms across the molecular function and biological process namespaces reveals key characteristics of these enzymes. (Fig 10)

**Molecular Function:** A significant portion of the identified GO terms falls under the molecular function namespace. Notably, several enzymes are associated with **binding activities**, including "carbohydrate binding" (GO:0030246), "ATP binding" (GO:0005524), "DNA binding" (GO:0003677), and "metal ion binding" (GO:0046872). This suggests that interactions with various molecules, including substrates, cofactors, and nucleic acids, are crucial for the function of these enzyme classes.

Furthermore, specific enzymatic activities are also represented, such as "glucan 1,4-alpha-glucosidase activity" (GO:0016759), "2-oxoglutarate-dependent dioxygenase activity" (GO:0016706), indicating the diverse catalytic roles of these enzymes.

The presence of "ATP-dependent activity" (GO:0140657) further emphasizes the role of ATP in the function of some of these enzymes.

**Biological Process:** The analysis also identified GO terms within the biological process namespace, providing a broader context for the cellular roles of these enzymes. The term "carbohydrate metabolic process" (GO:0005975) suggests the involvement of some enzyme classes in the synthesis or breakdown of carbohydrates, aligning with the "carbohydrate binding" and glycosidase/glycosyltransferase activities observed in the molecular function analysis. Additionally, the presence of "signal transduction" (GO:0007165) indicates that at least one of the enzyme classes may play a role in cellular signalling pathways.

Enzyme name	GO ID	Function	Namespace
GH36-type glycosyl hydrolase domain-containing protein [Pseudomonas aeruginosa]	GO:0030246	carbohydrate binding	molecular_function
GH36-type glycosyl hydrolase domain-containing protein [Pseudomonas aeruginosa]	GO:0005975	carbohydrate metabolic process	biological_process
GH36-type glycosyl hydrolase domain-containing protein [Pseudomonas aeruginosa]	GO:0004339	glucan 1,4-alpha-glucosidase activity	molecular_function
GH36-type glycosyl hydrolase domain-containing protein [Pseudomonas aeruginosa]	GO:0016757	glycosyltransferase activity	molecular_function
JmjC domain-containing protein [Pseudomonas aeruginosa]	GO:0016706	2-oxoglutarate-dependent dioxygenase activity	molecular_function
conjugal transfer nickase/helicase domain-containing protein, partial [Pseudomonas aeruginosa]	GO:0005524	ATP binding	molecular_function
conjugal transfer nickase/helicase domain-containing protein, partial [Pseudomonas aeruginosa]	GO:0003677	DNA binding	molecular_function
conjugal transfer nickase/helicase domain-containing protein, partial [Pseudomonas aeruginosa]	GO:0003678	DNA helicase activity	molecular_function
diguanylate cyclase domain-containing protein [Pseudomonas aeruginosa]	GO:0052621	diguanylate cyclase activity	molecular_function
diguanylate cyclase domain-containing protein [Pseudomonas aeruginosa]	GO:0005525	GTP binding	molecular_function
diguanylate cyclase domain-containing protein [Pseudomonas aeruginosa]	GO:0046872	metal ion binding	molecular_function
diguanylate cyclase domain-containing protein [Pseudomonas aeruginosa]	GO:0007165	signal transduction	biological_process
terminase large subunit domain-containing protein [Pseudomonas aeruginosa]	GO:0005524	ATP binding	molecular_function
terminase large subunit domain-containing protein [Pseudomonas aeruginosa]	GO:0140657	ATP-dependent activity	molecular_function
terminase large subunit domain-containing protein [Pseudomonas aeruginosa]	GO:0004519	endonuclease activity	molecular_function
terminase large subunit domain-containing protein, partial [Pseudomonas aeruginosa]	GO:0005524	ATP binding	molecular_function
terminase large subunit domain-containing protein, partial [Pseudomonas aeruginosa]	GO:0005524	ATP binding	molecular_function

Fig 10: - All mapped Go Id's of enzyme class

## String-DB Protein-Protein Interaction Network Analysis

A protein-protein interaction (PPI) network (Fig 10) was constructed using String-DB based on the list of proteins associated with the identified GO IDs

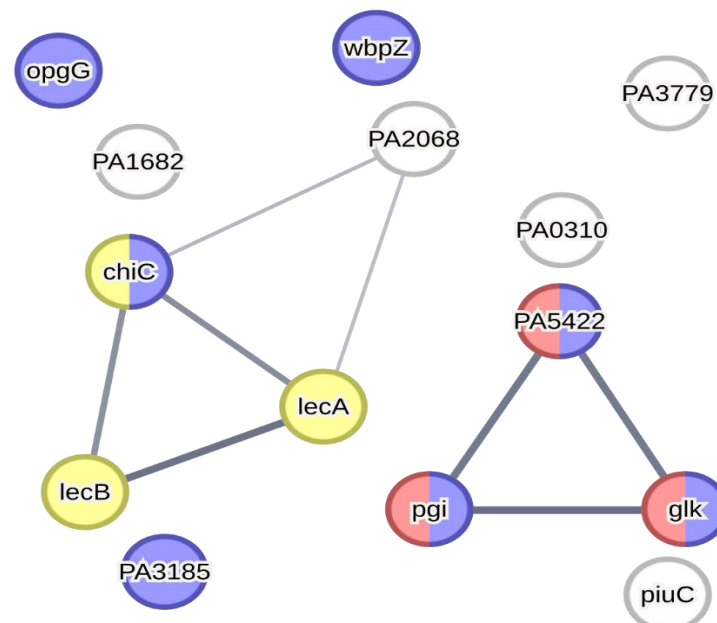


Fig 10: - Protein-Protein Interaction Network

- There are 14 nodes (Fig 10) in these networks represents the *Pseudomonas aeruginosa* proteins (identified by their gene names) associated with a specific GO term.
- Edges indicated known or predicted functional interactions between these proteins.

**Glycolysis** is represented by **orange** nodes. The network shows a distinct cluster of *pgi*, *glk*, *PA5422* connected by edges, indicating a tightly regulated and interacting set of proteins dedicated to this central metabolic pathway. This clustering suggests a high degree of functional coordination within the glycolytic machinery of *Pseudomonas aeruginosa*.

Proteins participating in **broader carbohydrate metabolic processes** are depicted as **blue** nodes. They exhibit connections within the network, suggesting a coordinated network involved in various aspects of carbohydrate metabolism beyond glycolysis. The interactions between these blue nodes may represent regulatory mechanisms or shared components within different carbohydrate-related pathways.

**Yellow** nodes represent proteins implicated in **quorum sensing**. The presence of interactions involving these yellow and blue nodes suggests potential links between quorum sensing mechanisms and the carbohydrate metabolic processes. This could indicate that quorum sensing, a cell-density-dependent communication system, influences or is influenced by the metabolic state of the bacterium.

## KEGG Pathway Analysis of Glycolysis/Gluconeogenesis

To further contextualize the functional roles of the identified proteins, we mapped three closely associated proteins from the String-DB network (Fig 10) – phosphoglucose isomerase (pgi, EC 5.3.1.9), a protein encoded by PA5422 (EC 5.1.3.15), and glucokinase (glk, EC 2.7.1.2) – onto the KEGG glycolysis/gluconeogenesis pathway (pae00010). (Fig 11)

The visualization of these proteins within the KEGG pathway (Fig 11) reveals their key positions and potential influence within central carbon metabolism.

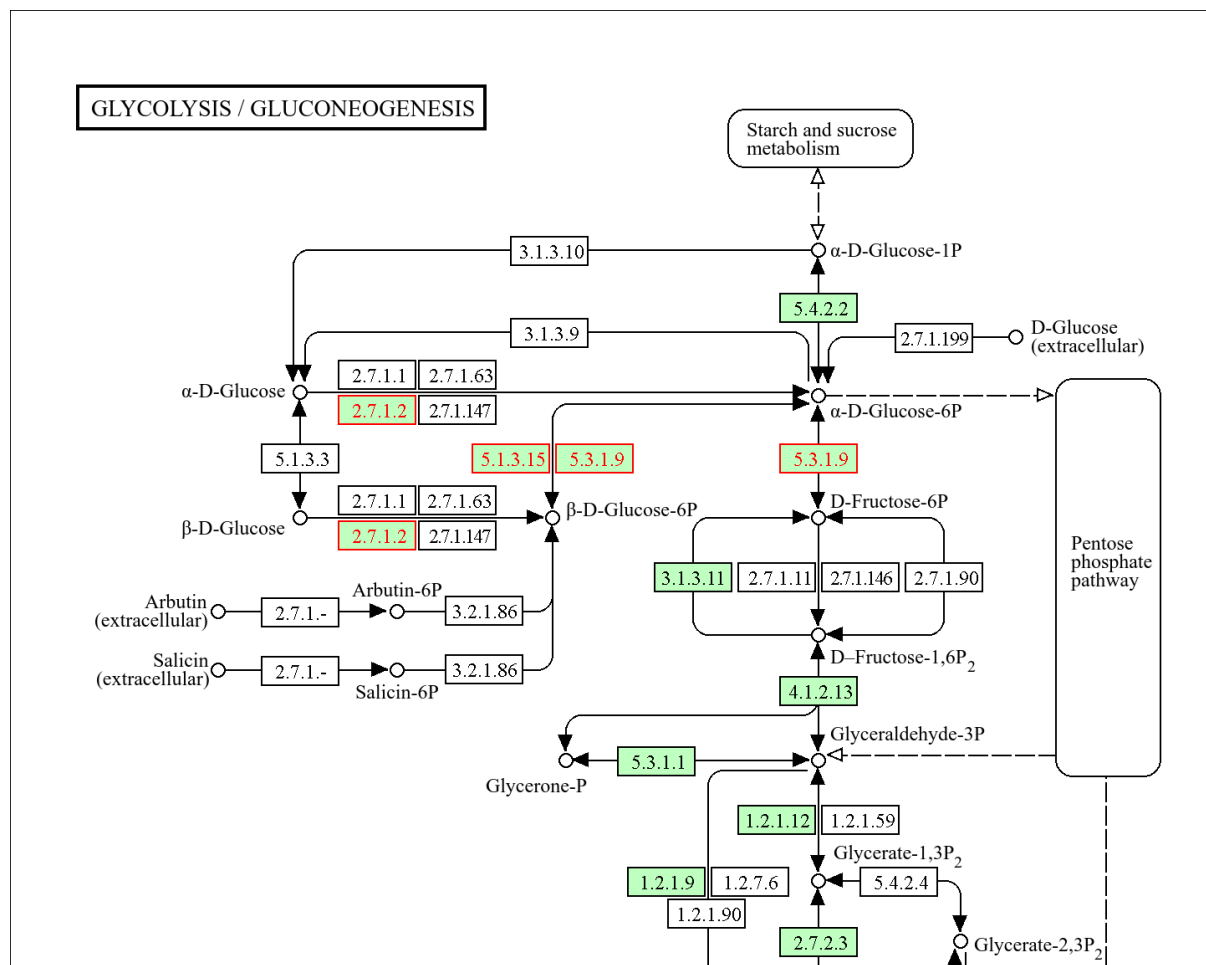


Fig 11: - KEGG pathway for Glycolysis

- **Phosphoglucose isomerase (pgi, EC 5.3.1.9):** As depicted in the pathway, PGI catalyses the reversible isomerization of glucose-6-phosphate to fructose-6-phosphate, a crucial step in both glycolysis and gluconeogenesis. Its central location highlights its role in directing the flow of carbon through these opposing metabolic routes.
- **PA5422 (EC 5.1.3.15):** In the context of the glycolysis/gluconeogenesis pathway, its potential phosphoglucomutase activity suggests a role in the conversion of

glucose-1-phosphate to glucose-6-phosphate, linking glycogen breakdown or other sugar metabolism pathways to the main glycolytic route.

- **Glucokinase (glk, EC 2.7.1.2):** Glucokinase catalyzes the phosphorylation of glucose to glucose-6-phosphate, the first committed step in glycolysis. Its presence at the entry point of the pathway underscores its importance in initiating glucose metabolism within *Pseudomonas aeruginosa*.

## Functional Annotation Results

By integrating the GO term assignments (derived from predicted enzyme classes), protein-protein interaction networks (generated via STRING-DB (Fig 10) for *Pseudomonas aeruginosa* proteins sharing GO terms), and KEGG pathway mapping (Fig 11), a multi-layered understanding of the predicted protein functions was achieved.

This integrated approach yielded the following key insights:

1. **Prioritizing Experimental Efforts:** Faced with many unannotated proteins, this approach helps prioritize those that are likely to be involved in key biological processes or have significant interaction partners, making research more focused and efficient.
2. **Drug Target Identification:** Understanding the precise roles of proteins within disease-related pathways can aid in the identification of potential drug targets. Knowing the protein's function, its interactions, and its position in a pathway is crucial for rational drug design.
3. **Understanding Disease Mechanisms:** By annotating proteins that are differentially expressed or mutated in disease states, we can gain a clearer understanding of the molecular mechanisms underlying these conditions.
4. **Comparative Genomics:** Applying this pipeline to different organisms can reveal conserved or divergent functional roles of proteins and pathways, providing insights into evolutionary biology and species-specific adaptations.



## Discussion

The findings of this study provide compelling evidence that physicochemical features can serve as reliable predictors of protein function in bacterial organisms. The superior performance of the Random Forest classifier, with its ensemble learning approach and ability to model complex non-linear relationships, underscores its suitability for this domain. The deliberate exclusion of primary sequence data marks a significant departure from conventional sequence-based function prediction methods, highlighting the inherent predictive power of structural and chemical properties in functional annotation tasks.

The successful integration of machine learning predictions with established biological knowledge through GO term mapping, PPI network analysis, and pathway mapping underscores the translational potential of this approach, particularly in scenarios involving the annotation of novel proteins or the functional characterization of proteins in poorly annotated organisms. The consistency observed between the predicted functional classes and their representation within known protein interaction networks and metabolic pathways provides strong support for the biological relevance and accuracy of the proposed methodology.

### Limitations:

- **Removal of Rare Classes:** A major limitation of this pipeline is the removal of protein classes with fewer than 25 instances. These rare classes, while statistically challenging for model training, can be biologically very important and may represent specialized functions unique to certain cellular processes or conditions.
- **Physicochemical Feature Limitations:** While informative, physicochemical features alone may not capture the full complexity of protein function, which is also influenced by tertiary structure, post-translational modifications, and cellular context.
- **GO Term Resolution:** The accuracy of GO term mapping depends on the available annotations for the predicted enzyme classes. Some predictions might map to broad or less specific GO terms, limiting the depth of functional insight.
- **STRING-DB and KEGG Coverage:** The coverage of protein interactions and pathway information in STRING-DB and KEGG for *Pseudomonas aeruginosa* is not exhaustive. Some predicted proteins might lack corresponding entries or detailed pathway information, limiting the scope of annotation.

## Future Work:

Several avenues for future research can build upon the findings of this study:

- **Expanding Feature Sets:** Including 3D coordinate information of the protein structure for model training is another promising avenue to explore, as it directly relates to protein function.
- **Addressing Class Imbalance:** Using bootstrapping techniques to address class imbalance should also be investigated.
- **Exploring Advanced Models:** Exploring the use of Graphical Neural Networks (GNNs) is also a promising direction.
- **Integrating Sequence Information:** Exploring hybrid approaches that combine these features with sequence-derived information

## **Conclusion**

In this project, we used machine learning to predict protein function based on their physical and chemical characteristics in *Pseudomonas aeruginosa*. We found that a Random Forest model worked best for this prediction. To understand the biological meaning of these predictions, we used tools to link the predicted functions to known Gene Ontology terms, explored how these proteins might interact with each other within the organism, and identified the biological pathways they likely participate in. This approach shows that we can gain valuable insights into protein function by combining machine learning with established biological knowledge.

# Appendices

1. Dataset - [Pseudomonas aeruginosa Protein Dataset](#) (Kaggle)
  
2. Websites –
  - <https://string-db.org/>
  - <https://www.genome.jp/kegg/pathway.html>
  
3. Code extract – <https://github.com/adrija-tih/Ideas-TIH-Internship-2025-Protein-Function-Prediction->
  
4. Document Link - <https://github.com/adrija-tih/Ideas-TIH-Internship-2025-Protein-Function-Prediction->