Problem Statement: Develop a model to predict the immunogenicity of novel antigens for vaccine design. **(PE5)**

1) Understanding the Problem Statement:

"Immunogenicity" refers to the ability of a substance to induce cellular and humoral immune responses.[1] In layman's terms, when a foreign body enters the human body, the ability of that body to provoke an immune response is referred to as "Immunogenicity". It is an essential factor during vaccine design.

Antigens, the foreign bodies that enter the human system, trigger an immune response when they are recognised as non-self or foreign. Immunogenicity is a trait or property of antigens. Few antigens are highly immunogenic while few may not trigger a strong immune reaction.

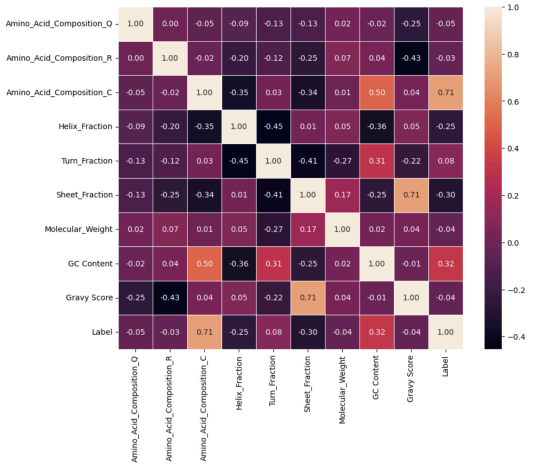
It is essential to determine the immunogenicity of protein-based therapeutics. Vaccines, when designed and experimented with, will still act as a foreign body triggering an unwanted immune response and nullifying the desired effect. To tackle this, the immunogenicity levels of the contents in a vaccine need to be maintained accordingly.

The problem statement aims to develop a model that can predict the immunogenicity of antigens for enhancing vaccine development. By automating this particular process of development, the process of vaccine development and implementation becomes easier in the initial stages.

2) Data Handling and Analysis:

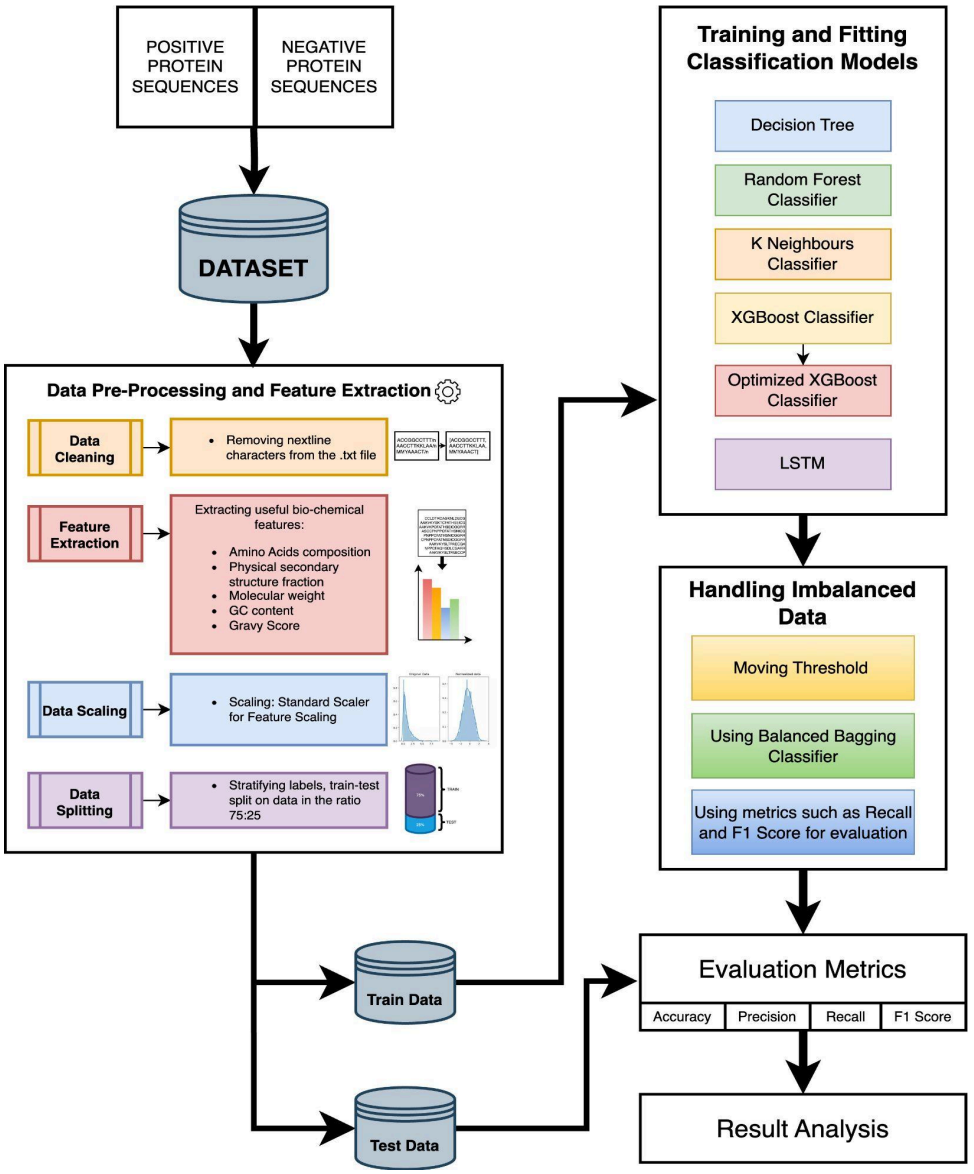
Data handling and analysis for the project involved a systematic approach to ensure dataset quality. After loading and inspecting raw data, cleaning procedures were applied. Feature extraction focused on identifying and transforming relevant data for model training, with consideration for feature scaling. Label 0 stands for Negative dataset which depicts that the protein sequence is not immunogenic whereas label 1 stands for positive data depicting immunogenic protein sequence.

The analysis phase included exploratory data analysis (EDA) using tools like Plotly and Matplotlib to gain insights into data distribution and relationships between features. Visualizations, including bar plots and correlation matrices, were used to identify trends and outliers.



A class imbalance in the dataset was observed. Techniques such as adjusting classification thresholds, optimizing hyperparameters, and implementing ensemble methods like the BalancedBaggingClassifier were explored to enhance model robustness.

3) Methodology:



The methodology started with the installation of necessary dependencies, encompassing libraries and packages essential for data manipulation, machine learning, and visualization. Following this, the raw data underwent preprocessing, involving data cleaning, feature extraction and normalization procedures. We extracted in total 9 bio-chemical features from the given amino acid sequences by applying appropriate mathematical formulas and logic. The features are mentioned below.

3.1) Protein Structure:

Protein structure is one of the factors that can affect the immune response in man.[2] There are 3 main types of Secondary Structures of proteins, namely:

1) Alpha Helix
2) Beta Sheets
3) Turn

The chemical structure (including amino acid sequence, glycosylation, and pegylation) can influence the incidence and level of antibody formation. [2]

Therefore, after thorough research and careful consideration, these three structural types have been taken into consideration. They act as individual features for the model to determine the immunogenicity of antigens.

3.2) GRAVY Score:

Grand average of hydropathicity index (GRAVY) is used to represent the hydrophobicity value of a peptide[3]. The relationship between hydrophobicity, hydrophilicity and immunogenicity is a proven fact.

Highly hydrophobic and highly hydrophilic polymers are not immunogenic. Moderate hydrophobicity as well as moderate hydrophilicity, and solubility in water favour immunogenicity (provided the molecular mass be at least 10,000 Da). For example, the solubilisation of zein (a hydrophobic insoluble maize protein) prior to immunization causes zein to become immunogenic. [5]

Therefore, after thorough consideration and research, it has been concluded that the GRAVY Score would be a key factor in determining the immunogenicity of antigens.

3.3) Amino Acid Composition:

A standard property for the determination of many factors and also immunogenicity is Amino Acid Composition. Often used as a parameter to analyse immune triggers in the human body, amino acid composition stands tall in its relevance when it comes to immunology.

For the better working of the model, only 3 amino acids have been considered. Namely, arginine, glutamine and cysteine.

Increasing evidence shows that dietary supplementation of specific amino acids to animals and humans with malnutrition and infectious disease enhances the immune status, thereby reducing morbidity and mortality. Arginine, glutamine and cysteine precursors are the best prototypes.[8]

Thus, from the above source and understanding, considering their importance and vitality in determining immune status, these three amino acids have been considered and used as 3 separate features for determining the final result.

3.4) GC Content:

GC content, or guanine-cytosine content, is the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine (G) or cytosine (C). GC Content has a strong correlation with

temperature. Previous studies consistently showed positive correlations between growth temperature and the GC contents of structural RNA genes.[6]
The relationship between temperature and immunogenicity has been proven by many scientific researchers in the past. The structural disorganization caused by high temperatures, may contribute to the reduction of immunogenicity.[7] This relation is a unique way to look at the immunogenicity of antigens. More GC content implies that the melting point of the protein sequence or antigen is higher, leading it to be a stable molecule and resistant to temperature changes. Even at high temperatures when there is a chance for structural disorganization to occur, molecules with high GC content (high melting point), will withstand the changes and effects.
GC content shapes amino acid composition to trade off the cost of amino acids with bases, which could be caused by energy efficiency.[11] As previously mentioned, amino acid composition is an essential factor in immunogenicity determination, and the fact that GC content is interconnected with amino acid composition makes it a relevant and important criterion to consider.
Thus, GC Content has been taken as a feature for the evaluation and determination of the immunogenicity of antigens in this model.

3.5) Physical Composition - Molecular Weight:

The physical composition of molecules and compounds has been a standard factor for the determination of many properties and classifications. Similarly, a physical factor i.e., molecular weight plays an important role in determining the immunogenicity of molecules.
An important factor affecting immunogenicity is antigen molecular weight.[9] The requirement for immunogenicity is high molecular weight. Small compounds (MW less than 1000), such as penicillin, progesterone and aspirin, as well as many moderately sized molecules (MW from 1000 to 6000), are not immunogenic. Most compounds with a molecular weight greater than 6000 are immunogenic. [10]
Considering the above scientifically proven claims, molecular weights of the protein sequences have been included as a feature in determining the immunogenicity of molecules or antigens.

Visualization played a crucial role in the exploratory phase, utilizing tools such as Plotly and Matplotlib to represent data distributions and feature relationships visually. A class imbalance in the dataset was observed, therefore we then performed Data Handling techniques. First, model selection followed, with a careful choice of machine learning models and the subsequent division of data into training and testing sets. The class imbalance was then addressed through various techniques. For instance, Technique 1 involved the exploration of moving thresholds. Based on that, hyperparameter optimization, and the implementation of an LSTM model for sequence-related data was performed for evaluation. Technique 2 leveraged the BalancedBaggingClassifier, enhancing the model's performance through additional bagging. The final stages focused on the evaluation and analysis of results, employing metrics such as accuracy, precision, recall, and F1-score to take care of the data imbalance. The outcomes were visually presented, allowing for a comprehensive summary and interpretation of the project's

findings. Throughout the project, each step was thoroughly documented, ensuring a systematic and well-documented approach.

4) Mathematical and Logical Reasoning:

4.1) Probability of Immunogenicity Determination:

The final probability of immunogenicity is determined through a function that is often used for Machine Learning for classification. This function deals with two modes: Multinomial Mode and One-vs-Rest Mode.

One-vs-Rest Mode is used for binary classifications. Since our model leverages the use of Binary Classification, this mode has been implemented.

For a given input vector X, the function is applied to the linear combination of the input features and class-specific parameters:

$$P(Y_i = 1|X) = 1/ (1 + e^{-[X.W_i + b_i]})$$

Here: W_i and b_i are the weight vector and bias for class i,

e is the base of the natural logarithm

This process is repeated for each class, resulting in a probability for each class. The probabilities across all classes for a given sample are then normalized to sum to 1.

4.2) The GRAVY (Grand Average of Hydropathy) score:

This score is calculated by the sum of hydropathy values of all amino acids divided by the protein length

GRAVY Score = ∑(Hydropathy values of Amino Acids) / (Length of protein sequence)

Hydropathy Values of Amino Acids are calculated through various scales. The most common one is the Kyte-Doolittle Hydropathy.

The Kyte-Doolittle hydropathy scale is based on the free energy change when an amino acid is transferred from a hydrophobic environment to water. The scale is calculated using a sliding window of amino acids, and the formula for calculating the hydropathy index is as follows:

$$H_i = 1/n \sum_{j=1}^{n} h_j$$

H_i is the hydropathy index for i-th amino acid.

N is the window size

H_j is the hydropathy value of the j-th amino acid in the window.

The Kyte-Doolittle scale typically uses a window size of 7 amino acids. The hydropathy values for each amino acid are predefined and summed up over the window size to calculate the hydropathy index for each position in the protein sequence.

4.3) GC Content:

GC content is usually calculated as a percentage value and is sometimes called G+C ratio or GC-ratio. GC-content percentage is calculated:

$$[Count(G) + Count(C)]/ [Count(A) + Count(T) + Count(G) + Count(C)] * 100$$

4.4) Amino Acid Composition:

Calculated by counting the values of 3 amino acids: R,C,Q.
These three amino acids are influential in determining immunogenicity, hence the count of these three amino acids has been considered.

4.5) Protein Structure and Fractions:

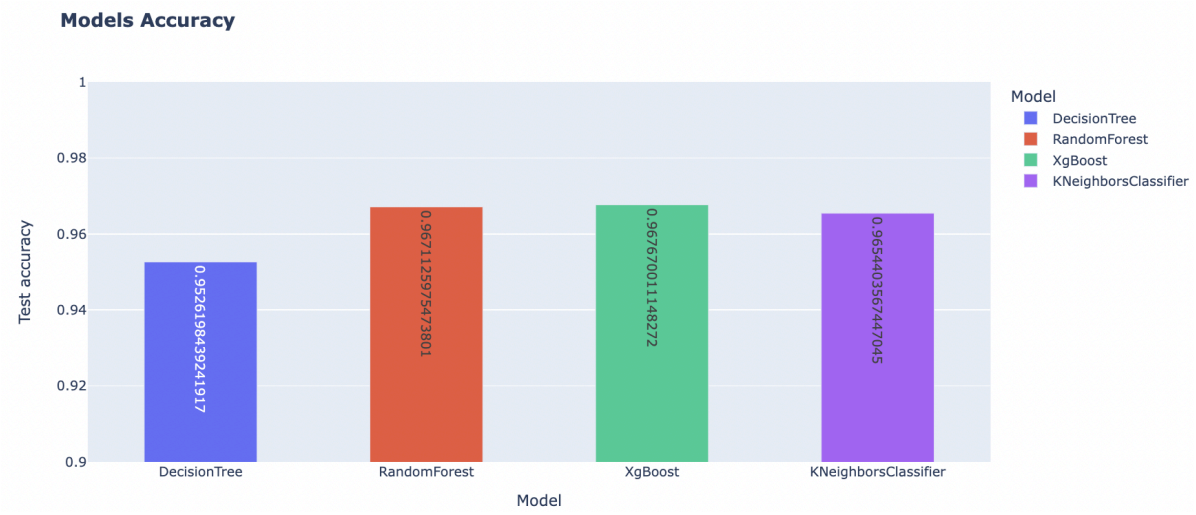
Amino acids in helix: V, I, Y, F, W, L.
Amino acids in turn: N, P, G, S.
Amino acids in sheet: E, M, A, L.

Helix Fraction:
$Total\ number\ of\ amino\ acids\ belonging\ to\ helix\ /\ Total\ Number\ of\ amino\ acids$
Sheet Fraction:
$Total\ number\ of\ amino\ acids\ belonging\ to\ sheet\ /\ Total\ Number\ of\ amino\ acids$
Turn Fraction:
$Total\ number\ of\ amino\ acids\ belonging\ to\ turn\ /\ Total\ Number\ of\ amino\ acids$

5) Evaluation Metrics

Precision, recall and F1-Score were used as metrics for evaluation. These metrics have been used for the following reasons:
i) Dependability
ii) Handling data imbalance
iii) Easier understanding of the model
To deal with the imbalanced dataset, we incorporated 2 techniques; moving threshold value, and using a Balanced bagging classifier.

4 different Machine Learning Models and their accuracies on the dataset are displayed below:



These accuracies display that XGBoost is the most efficient and highly accurate model out of all. Hence, it has been selected for further Threshold optimisation.

The classification evaluation report after moving the threshold of the model is given below:
XGBoost with optimized parameters (Optimized_XGBoost):

Test Accuracy: 0.9930

	precision	recall	f1-score	support
0	1.00	1.00	1.00	15966
1	0.97	0.96	0.97	1793
accuracy			0.99	17759
macro avg	0.98	0.98	0.98	17759
weighted avg	0.99	0.99	0.99	17759

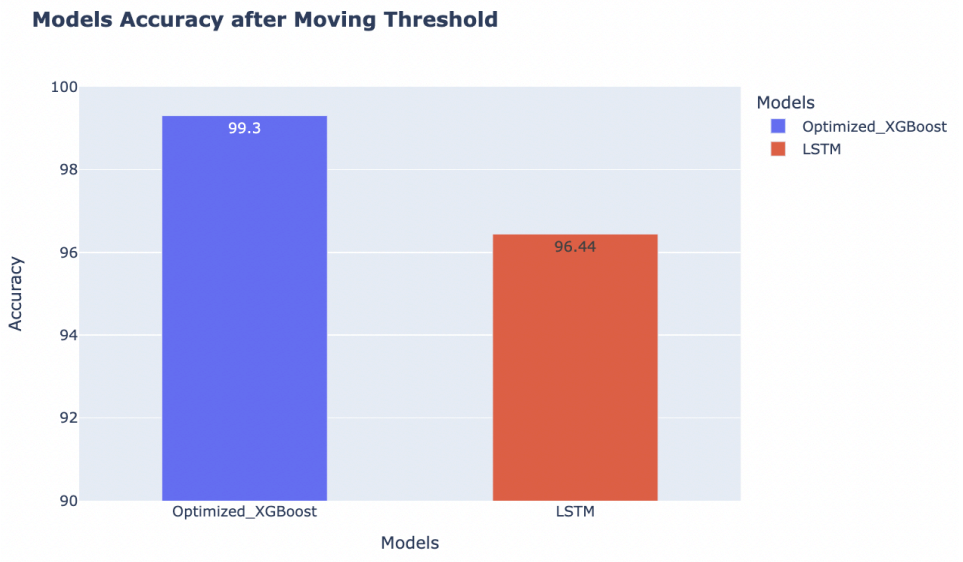
LSTM:

Test Accuracy: 0.9644

	precision	recall	f1-score	support
0	0.99	0.97	0.98	15966
1	0.79	0.89	0.83	1793
accuracy			0.96	17759
macro avg	0.89	0.93	0.91	17759
weighted avg	0.97	0.96	0.97	17759

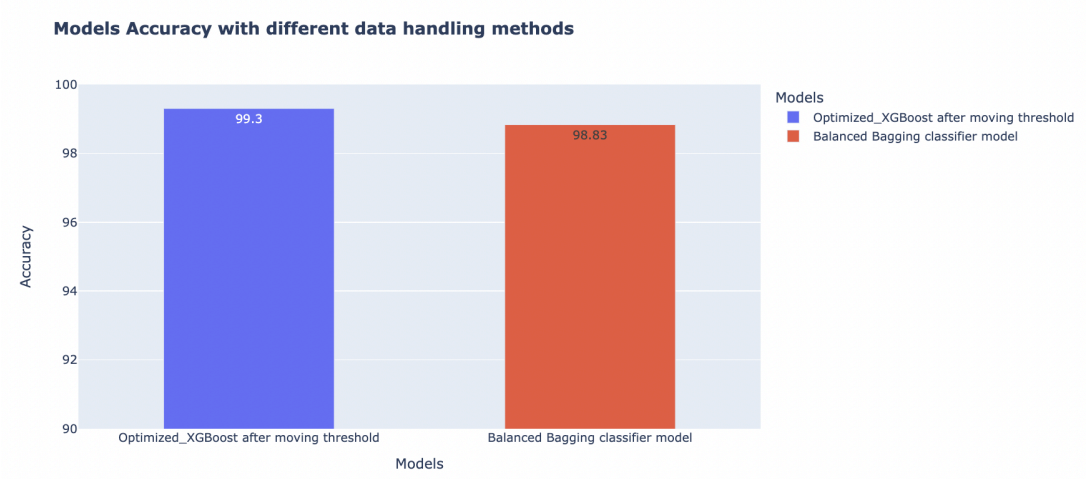
It can be observed that optimized XGBoost performed better than LSTM.

The below graph compares the accuracy of these models.



Apart from moving threshold, a Balanced bagging classifier was also used as a technique to deal with imbalanced datasets.

The evaluation graph and the classification report of this model are shown below.




```
Test Accuracy: 0.9883
              precision    recall  f1-score   support

           0       0.99      1.00      0.99     15966
           1       0.97      0.92      0.94      1793

    accuracy                           0.99     17759
   macro avg       0.98      0.96      0.97     17759
weighted avg       0.99      0.99      0.99     17759
```

Overall, moving threshold was seen as a better approach to deal with the imbalanced dataset, and the Optimized XGBoost classifier achieved the highest accuracy of 99.3%.

6) Benchmarking with other models:

To compare and analyse the key factors of model preparation, three models have been considered to benchmark our model with. 4 factors are used to benchmark the models. The factors used and detailed benchmarking analysis is done in the table below.

References:
Model 1: Bacterial Immunogenicity Prediction by Machine Learning Methods
Model 2: DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity
Model 3: DeepBCE: Evaluation of deep learning models for identification of immunogenic B-cell epitopes

Factors	**Our Model**	**Model 1**	**Model 2**	**Model 3**
Machine Learning Approach	Explore Decision Tree Classifier, Random Tree Classifier, K-Nearest Neighbors Classifier, and XGBoost (Extreme Gradient Boosting) optimized using random search CV. Balanced Bagging Classifier has been utilised to address the problem of imbalanced data. Employs Deep Learning using LSTM for data mining of immunogenicity.	Utilizes SVM, RFC, and LSTM for predicting B-Cell epitopes. Focuses on feature engineering with PACVF features	Explores SVM, RFC, and various deep learning models considering factors that determine immunogenicity. Features include PACVF, and models are optimized using randomized search CV.	Incorporates SVM, RFC, GRU, and ConvNN for B-Cell epitope prediction, with specific details on architecture and hyperparameters for each model.
Data Used	Data of 17759 protein sequences. Out of which, 15966 rows of protein sequences are immunogenicity and 1973 rows of non-immunogenic protein sequences have been utilised.	The dataset used consists of 317 known bacterial immunogens and 317 bacterial non-immunogens.	Analyzed >9000 tested immunogenicity molecular assays from the Immune Epitope Database, IEDB database (13 August 2020). Restricted this dataset to peptides with metadata that matched the following keywords: (i) linear epitope, (ii) T-cell assay, (iii) MHC class I, (iv) human and (v) any disease.	A total of 30552 protein samples was obtained to form a benchmark dataset including 11834 pBCEs and 18722 nBCEs.
Factors considered	Protein Structure, Physical Composition and Molecular	Unique Epittotes, Protein Aggregates,	Amino Acid Sequences, Structural Modifications.	Antigenic Determinants, Protein Structure,

that are related to Immunogenicity	Weight, GRAVY Score and Hydrophobicity/Hydrophilicity, Amino Acid Composition, GC Content	Impurities of Antigens and Proteins.	Impurity Control, Conformational Changes	Host-Related Factors, Product-Based Impurities.
Accuracy and Robustness	The XGBoost model with optimized parameters gave the highest accuracy of approx. 99.3% while LSTM gave 96.44% Balanced Bagging Classifier was also used which showed an accuracy of 98.83%	Doesn't provide specific accuracy metrics. Details on the robustness are not explicitly mentioned.	Highlights ConvNN as the best-performing model with the highest mAP and AUC scores, suggesting good accuracy.	ConvNN-based model achieves the highest accuracy and MCC values, indicating superior performance and stability.

7) Future Scope

7.1) 3D Model Utilisation:

These proteins can be visualised in 3-dimensionality space. This can be achieved through AlphaFold and AlphaFold2. AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. The importance of 3D structure from a biological and chemical point of view is immense and is greatly appreciated by researchers and scientists all over the world.

The GPU consumption of this conversion and analysis is beyond a typical computer. Due to these GPU constraints and the complexity of proteins, the feature has not been included.

7.2) Mathematical Upscaling:

More thorough mathematical involvement can be done for the measurement of immunogenicity and several other factors. This would help achieve a statistical point of view for a model, aiding and assisting for easier improvements in the future.

References:

1) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4811736/#:~:text=The%20term%20immunogenicity%20refers%20to,response%20to%20the%20given%20substance.
2) https://link.springer.com/content/pdf/10.1023/B:PHAM.0000029275.41323.a6.pdf
3) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3734225/#:~:text=Grand%20average%20of%20hydropathicity%20index,divided%20by%20the%20sequence%20length.

4) https://karger.com/Article/FullText/508903#:~:text=immunogenicity
5) https://www.tandfonline.com/doi/abs/10.3109/08820139209069401
6) https://www.researchgate.net/publication/358473407_A_positive_correlation_between_GC_content_and_growth_temperature_in_prokaryotes
7) https://link.springer.com/article/10.1007/BF02436044
8) https://pubmed.ncbi.nlm.nih.gov/17403271/#:~:text=Increasing%20evidence%20shows%20that%20dietary,precursors%20are%20the%20best%20prototypes.
9) https://sites.ualberta.ca/~pletendr/tm-modules/immunology/70imm-term.html
10) https://www.thermofisher.com/in/en/home/life-science/antibodies/antibodies-learning-center/antibodies-resource-library/antibody-methods/antibody-production-immunogen-preparation.html#:~:text=Properties%20determining%20immunogenicity&text=The%20second%20requirement%20for%20immunogenicity,6000
11) https://www.researchgate.net/publication/329482168_The_GC_Content_as_a_Main_Factor_Shaping_the_Amino_Acid_Usage_During_Bacterial_Evolution_Process