Discontinuous Epitope Feature Extraction, and Analysis

IMMEDIATE

April 5, 2023

Yatindra Indoria, Srishty Gandhi, Devendra Palod (20CS30060,20CS3005,20CS100)

1. INTRODUCTION

Epitopes are specific regions on the surface of a molecule, such as a protein or a virus, that are recognized and bound by the receptors of immune cells, called antibodies. They are crucial for initiating an immune response and can be used as targets for developing vaccines or diagnostic tools. Epitopes can be linear or conformational, depending on their structure and arrangement, and their identification and characterization are essential for understanding the immune response to various pathogens and developing effective interventions.

We analyze a set of aggregation-based features, extracted base protein agnostically from epitopes over four different datasets. The goal is to quantify how effective the extracted features are in terms of understanding epitopes. What is it, that makes an epitope sequence different?

2. DISEASE ANALYSIS

A. Dengue

Dengue fever is a viral disease caused by the dengue virus, which is transmitted to humans through the bite of infected Aedes mosquitoes. Epitopes are small segments of a protein that are recognized by the immune system and can trigger an immune response. Epitopes are important for the development of vaccines and diagnostic tests for infectious diseases, including dengue fever. The dengue virus has four different serotypes, and infection with one serotype does not provide immunity to the other serotypes. This means that people can be infected with dengue fever multiple times, and each infection can be more severe than the previous one. There are several approaches to developing a vaccine for dengue fever, including using whole inactivated virus, live attenuated virus, or recombinant proteins. One challenge in developing a dengue vaccine is that the immune response to the virus can be complex, and a vaccine that triggers an immune response to one serotype may not provide protection against the other serotypes.

Researchers have identified several epitopes in the dengue virus that are recognized by the immune system, and these epitopes are being studied as potential targets for vaccines and diagnostic tests. Some researchers are also studying the use of computer modeling to identify new epitopes that could be targeted by a dengue vaccine.

Overall, dengue fever is a serious and complex disease, and developing an effective vaccine remains a challenge. However, researchers continue to make progress in understanding the virus and developing new approaches to prevent and treat dengue fever.

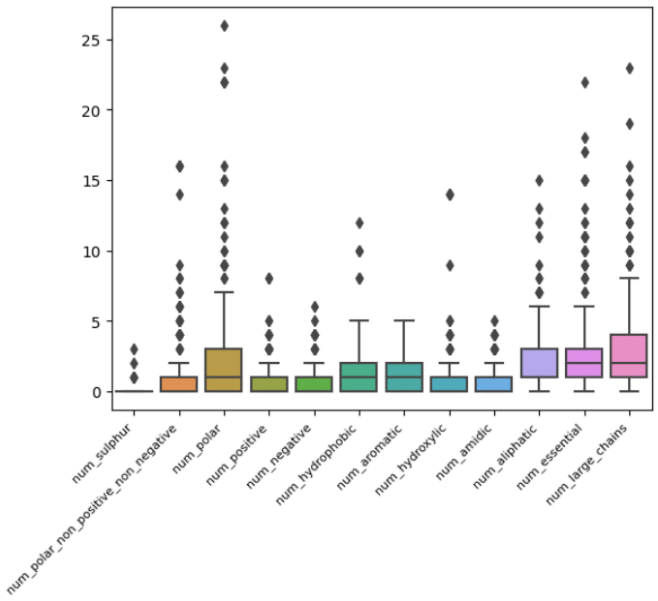

Fig. 1. Box plot for the disease Dengue

B. Malaria

Malaria is a life-threatening disease caused by the Plasmodium parasite, which is transmitted to humans through the bites of infected Anopheles mosquitoes. There has been significant research into the development of vaccines for malaria, with a number of potential vaccine candidates in various stages of development. One approach is to target specific epitopes on the Plasmodium parasite, such as the circumsporozoite protein (CSP), which is a major surface protein on the sporozoite stage of the parasite. CSP contains several B cell epitopes, which can be used to stimulate the production of antibodies that can neutralize the sporozoite stage of the parasite.

Other potential vaccine targets include the merozoite surface protein (MSP) and the apical membrane antigen 1 (AMA1),

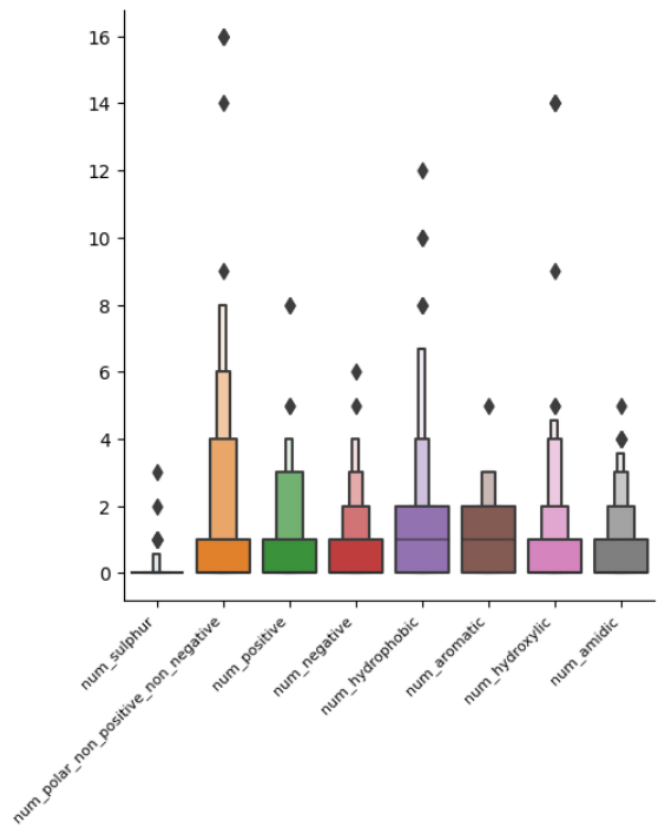

Fig. 2. Boxen plot for the disease Dengue

which are also major surface proteins on the merozoite stage of the parasite. These proteins contain multiple B and T cell epitopes, which can be used to stimulate both antibody and cellular immune responses.

In addition to vaccine development, the identification of epitopes can also be used for diagnostic purposes, such as the development of rapid diagnostic tests that can detect specific antibodies against the Plasmodium parasite in the blood of infected individuals.

C. Chikungunya

Chikungunya is a viral disease that is transmitted to humans by infected mosquitoes. The disease is caused by the chikungunya virus (CHIKV), which belongs to the alphavirus genus in the Togaviridae family. In the case of chikungunya, researchers have identified several epitopes that could be used in the development of a vaccine. One study, for example, identified a potential epitope on the E1 protein of the chikungunya virus that was able to induce a strong immune response in mice. Another study identified several potential epitopes on the E2 protein of the virus that could be used to develop a diagnostic test.

Overall, the identification and characterization of epitopes in the chikungunya virus is an important area of research that could lead to the development of new vaccines, treatments, and diagnostic tools for this disease.

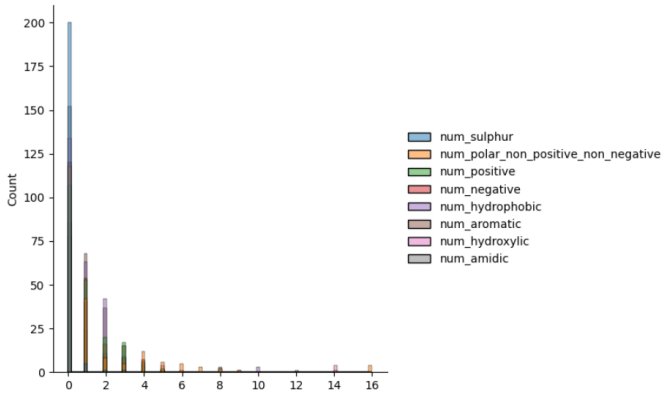

Fig. 3. count plot for the disease Dengue

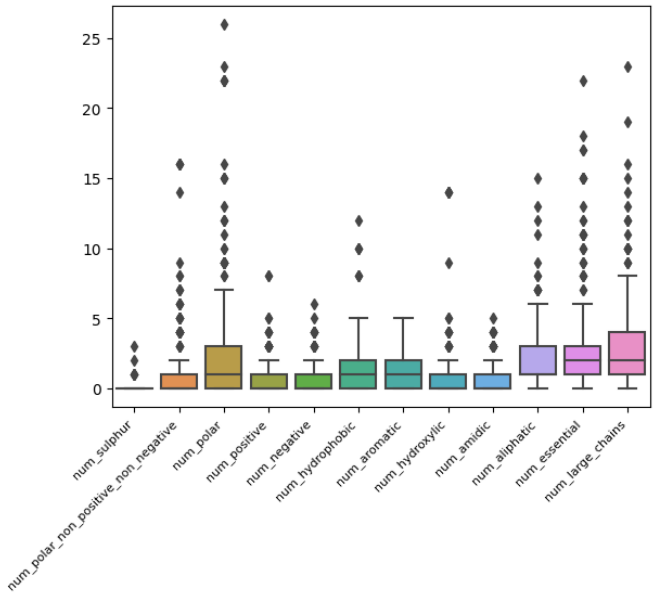

Fig. 4. Box plot for the disease Malaria

3. LITERATURE SURVEY

A. Literature 1

A.1. Overview

The development of reliable epitope prediction tools is a critical task in immunology and has significant implications for understanding disease pathogenesis, designing vaccines, and developing immune-based cancer therapies. However, experimental methods for epitope mapping are often time-consuming, expensive, and limited by technical challenges. To overcome these limitations, computational methods have been developed, and machine learning approaches have emerged as one of the most effective strategies.

A.2. Study

The development of reliable epitope prediction tools is not feasible in the absence of high quality data sets. Unfortunately, most of the existing epitope benchmark data sets are comprised of epitope sequences that share high degree of similarity with other peptide sequences in the same data set. We demonstrate the pitfalls of these commonly used data sets for evaluating the performance of machine learning approaches to epitope pre-

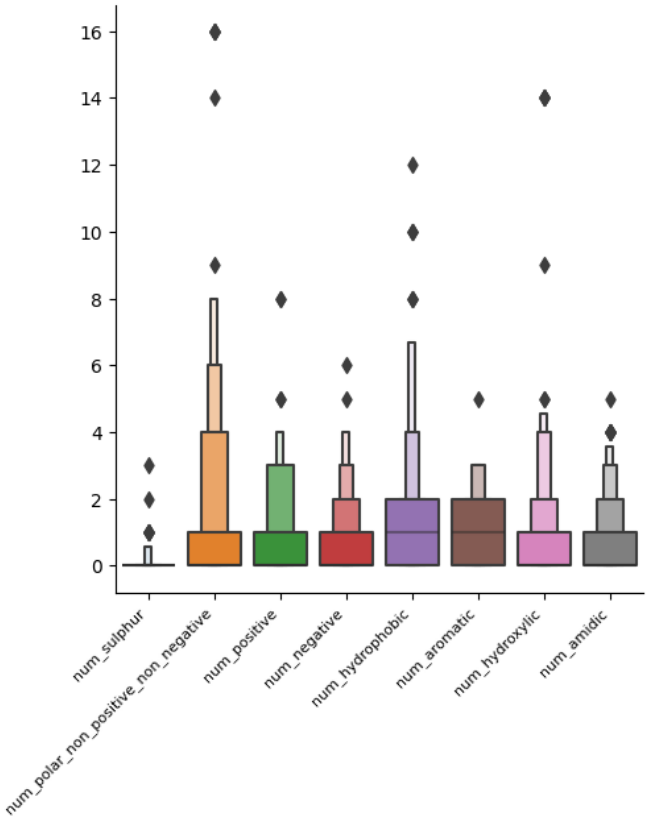

Fig. 5. Boxen plot for the disease Malaria

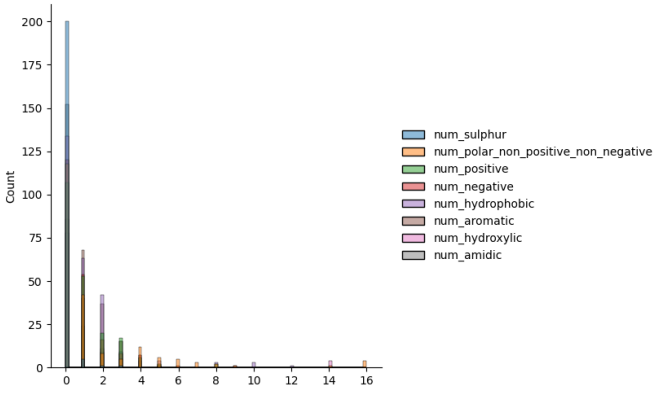

Fig. 6. count plot for the disease Malaria

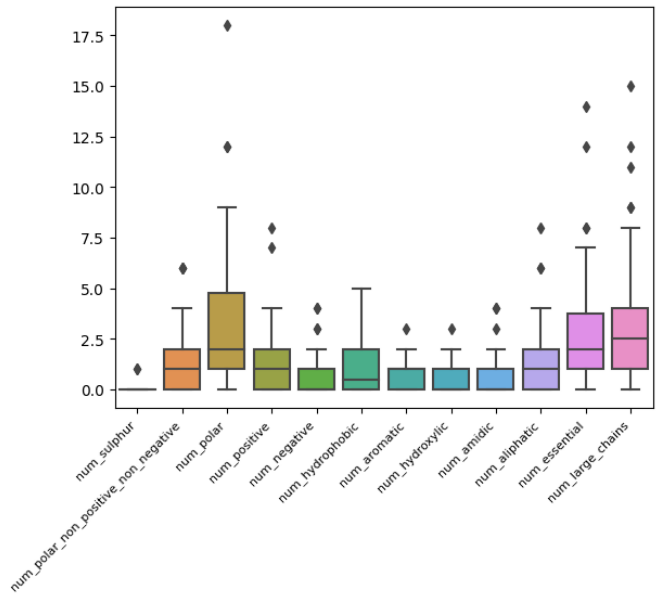

Fig. 7. Box plot for the disease Chikungunya

diction. Finally, we propose a similarity reduction procedure that is more stringent than currently used similarity reduction methods.

Two broad categories of prediction tools are presented: residue-based predictors and epitope-based predictors. Residue-based prediction methods assign binary labels to each individual residue in the input sequence, and each group of neighboring residues with predicted positive labels defines a variable length predicted linear B-cell epitope. Epitope-based predictors, on the other hand, take as input a protein sequence and an epitope length and apply a sliding window to extract peptides that are then passed to a neural network classifier.

Recently, a study by Xue et al. (2020) proposed several machine learning-based methods for epitope prediction. One of the significant challenges in predicting epitopes is the high degree of sequence diversity among different pathogens. To address this issue, the authors introduced two methods, BCPred and FBCPred, for predicting linear B-cell epitopes and flexible length linear B-cell epitopes, respectively, using string kernel-based support vector machine (SVM) classifiers. Both methods demonstrated high predictive performance on independent datasets, with FBCPred outperforming BCPred due to its ability to handle epitopes with flexible lengths. Several machine learning methods have been tested for linear B-cell epitope prediction, including BepiPred, ABCPred, and methods by S¨ollner and Mayer and Chen et al. The review notes that the performance of existing methods is only marginally better than random and presents the limitations of current epitope-based prediction tools, where the user is forced to select one of the available epitope lengths.

In the method of S¨ollner and Mayer (2006), each epitope is represented using a set of 1487 features extracted from a variety of propensity scales, neighborhood matrices, and respective probability and likelihood values. Of two machine learning methods tested, decision trees and a nearest-neighbor method combined with feature selection, the latter was reported. Chen et al. (2007) observed that certain amino acid pairs (AAPs) tend to occur more frequently in B-cell epitopes than in non-epitope peptides. Using an AAP propensity scale based on this observation, in combination with a support vector machine (SVM) classifier, they reported prediction accuracy of 71% on a data set of 872 B-cell epitopes and 872 non-B-cell epitopes.

We also utilized SVM with a RBF, and String kernel in our method, and that wasn't devoid of challenges. Although the performance of SVM-based classifiers largely depends on the selection of the kernel function, there are no theoretical foundations for choosing good kernel functions in a data-dependent way.

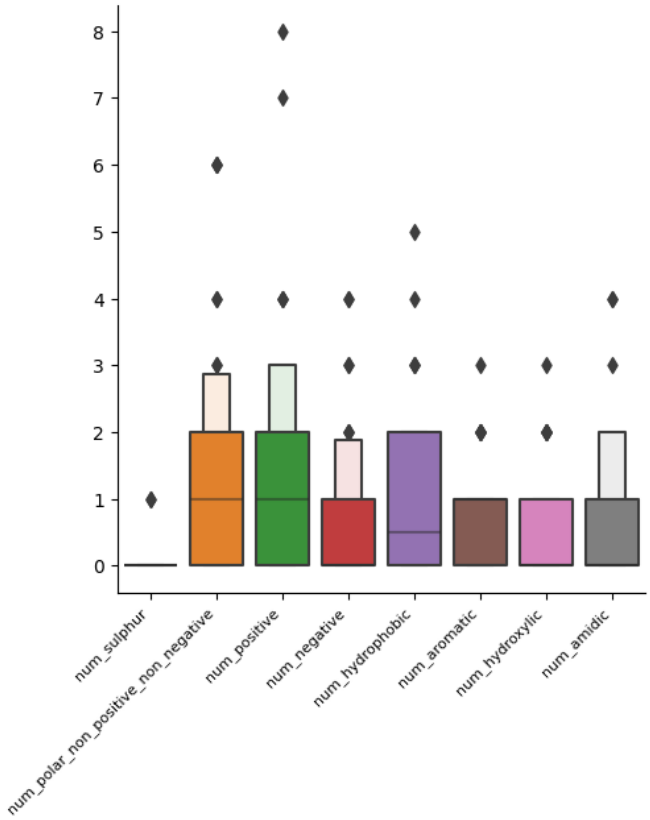

Fig. 8. Boxen plot for the disease Chikungunya

4. METHODOLOGY AND EXPERIMENTS

The work basically deals with the features and properties of epitope sequences that make them distinct from just another random sequence. We extract a custom 20 features from the epitopes involving the number of Sulphur atoms, and the number of positively charged, or negatively charged residues, among others. We also extract aggregate-based features from the AAIndex dataset, made available by genome.jp. We create a custom tool to read the epitope sequences and calculate the custom features as well as features corresponding to the AAIndex, and save them as a CSV file (and the tool comes with the functionality to download and parse the AAIndex data into a JSON (one-time thing)). The tool can read epitope sequences from CSV, XLS, TXT, and can scrape data off of the web as per requirement. The tool is also capable of generating random epitope sequences on existing proteins, as well as creating a synthetic protein sequence if none provided.

We then analyze the extracted features. We evaluated the correlations between the features and applied appropriate levels of dimensionality reductions to them (for both, the custom features, as well as the AAIndex features). We analyze, how good these features are before, and after the application of dimensionality reduction, at figuring out a totally random epitope sequence, from an actual epitope sequence.

We evaluate the accuracy of a binary classification task using the features we extract for the given epitopes over the four datasets, mixed with fake samples generated randomly. We use four different classifiers to stake the claims. The accuracies are averaged over a minimum of 5 runs, we ran the algorithms with overall data normalization, no normalization, and separate

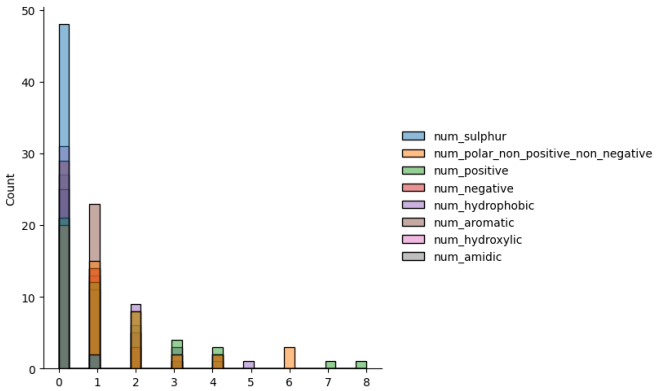

Fig. 9. count plot for the disease Chikungunya

normalization for fake and real data. The regressors used for classification

A. Dataset description

The first dataset used in this study consists of a collection of epitopes of B-cells. The dataset contains unique epitopes. Each epitope is represented as a string of amino acids and their occurrence index, comma seperated. Secondly, we have applied our analysis to make predictions on three more datasets. The dataset contains 6.6k entries, which shrink to 6.2k after data cleaning.

We use malaria, chikungunya, and dengue datasets. The datasets consist of a lot of metadata of the epitope sequences, including their IDs. The datasets for all three of the diseases contain varying numbers of entries (from 50 to 400).

B. Feature description

The number of polar, positive, or negative residues: we use the same to get a better idea of the interactions between 2 residues. Hence only the epitope residues need to be considered.

Hydroxylic, Amidic, Sulpur: Hydroxylic groups, Amidic groups not only show acidic/basic behaviour, but will also have the ability of forming H-Bond, and hence influence the bonding site. Similarly, 'C' can form S-S bonds, and hence is important
Hydrophilicity is the measure of how water-soluble a molecule or amino acid is. In epitope prediction, hydrophilicity is important as it helps in identifying antigenic regions of a protein that are more likely to be exposed to the aqueous environment and therefore potentially more accessible to antibodies.
Approximate volume refers to the size of the amino acid residue, and is important in epitope prediction as it affects the accessibility of the residue to antibody recognition. Smaller residues are more likely to be found on the surface of proteins, and therefore more likely to be antigenic.
Extinction coefficient is a measure of how much light a protein or amino acid absorbs at a particular wavelength. It is important in epitope prediction as it helps in the identification of protein regions that are likely to be exposed to the environment and potentially more accessible to antibodies.
Molecular mass is the sum of the masses of all the atoms in a molecule or amino acid. It is important in epitope prediction as it helps in the identification of protein regions that are likely to be exposed to the environment and potentially more accessible to antibodies.
Amino acid count refers to the number of amino acids in a

protein sequence. It is important in epitope prediction as it helps in the identification of potentially antigenic regions of the protein.
The amino acid percentage is the percentage of a particular amino acid in the entire protein sequence. It is important in epitope prediction as it helps in the identification of potentially antigenic regions of the protein.
Surface accessibility refers to the degree to which an amino acid residue is exposed to the environment. It is important in epitope prediction as it helps in the identification of potentially antigenic regions of the protein that are more likely to be exposed to antibodies.

Polarity is the measure of the degree to which an amino acid residue is hydrophilic or hydrophobic. It is important in epitope prediction as it helps in the identification of potentially antigenic regions of the protein that are more likely to be exposed to the environment and therefore more accessible to antibodies.
GRAVY score can be calculated as the sum of the hydropathy values for all the amino acids in a protein divided by the total number of residues in it.

C. Exploratory Analysis

Exploratory data analysis was conducted to gain insights into the distribution and characteristics of the dataset. Descriptive statistics were calculated for each variable, and visualizations were created to summarize the data. Some of the visualizations are attached below.

We created heatmaps for visualization correlation between the different features in consideration. We also visualized the ranges our features took, which was important to check the sanity of the random data being generated, and to verify if the algorithm was giving reasonable results.

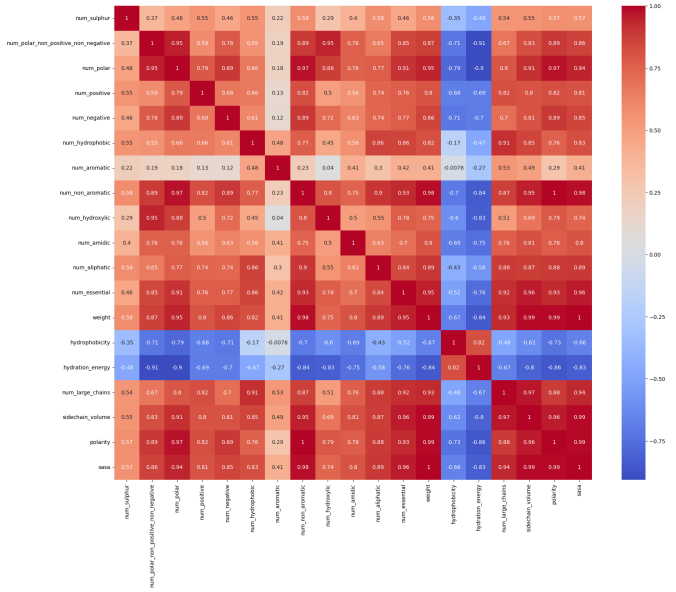

Fig. 10. heat map for the disease Dengue

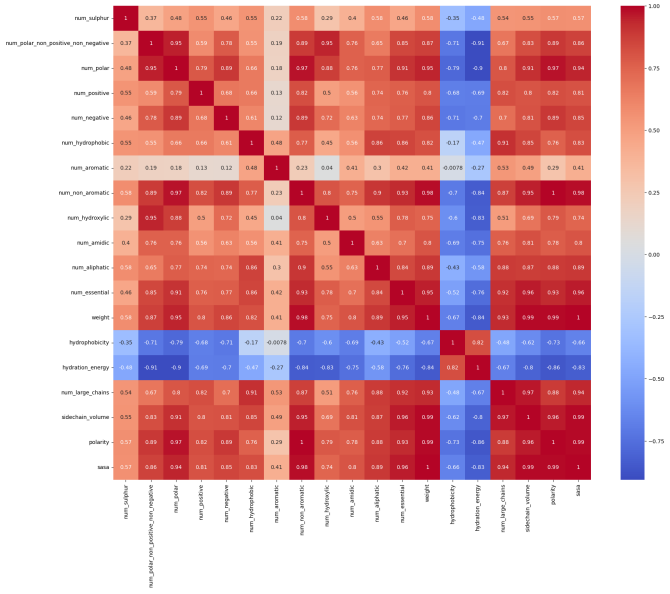

Fig. 11. heat map for the disease malaria

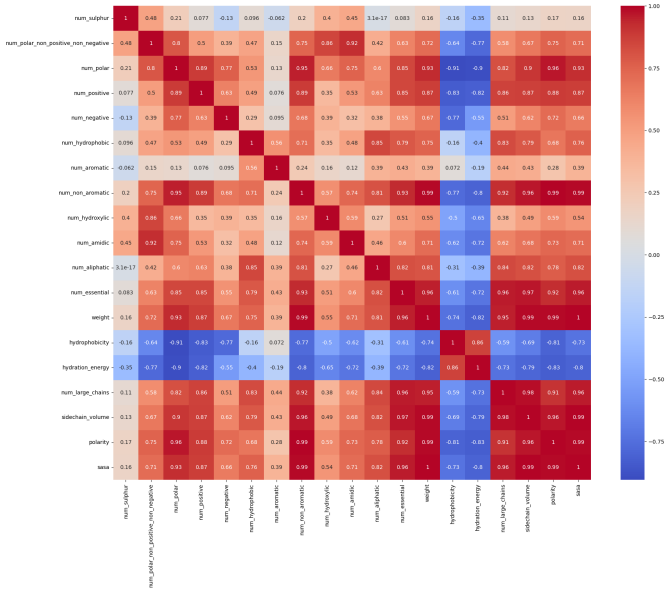

Fig. 12. heat map for the disease Chikungunya

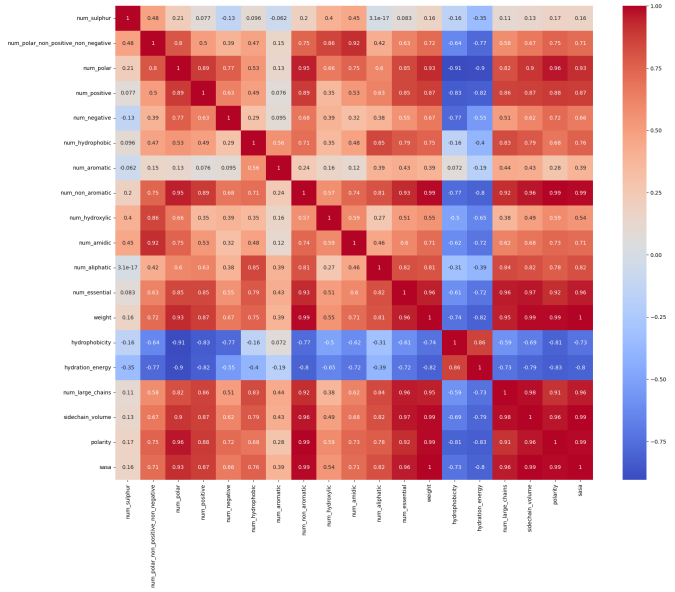

Fig. 13. heat map for the disease Chikungunya

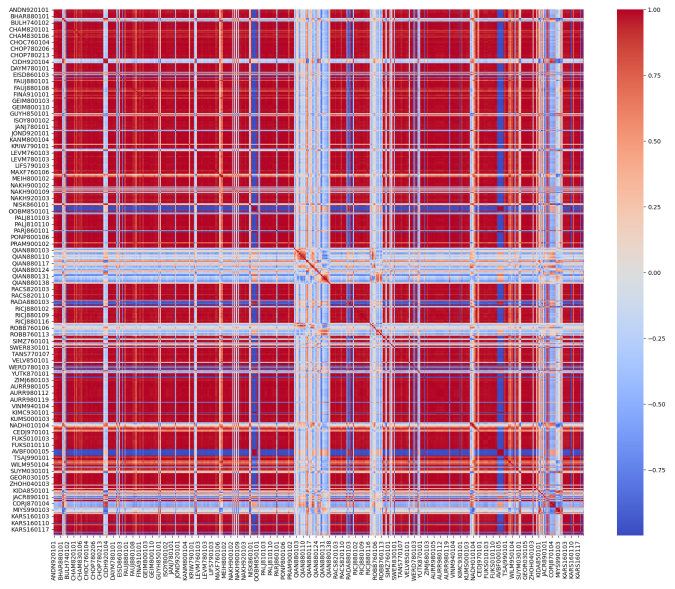

Fig. 14. heat map of AA index

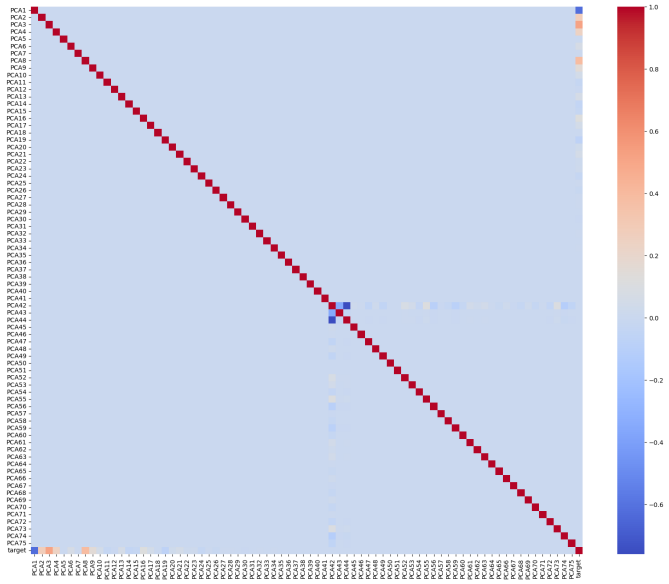

Fig. 15. heat map of AA index after PCA, final cols = 75

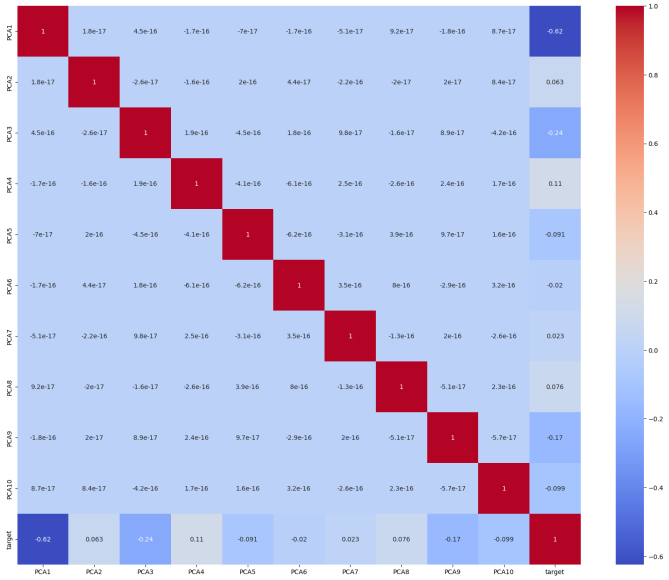

Fig. 16. Heat map(Dengue, PCA = 7)

5. RESULTS

We are attaching the table of accuracies obtained after PCA, and normalization over the Dengue dataset. Plots are given for all the datasets analyzed. We have obtained 80%+ accuracy in almost all runs. In some even more than that, but those results aren't presented here, as that was due to very low quality of random data generated (the random data when not properly curated, can lead to the generation of sequences which are impractical).

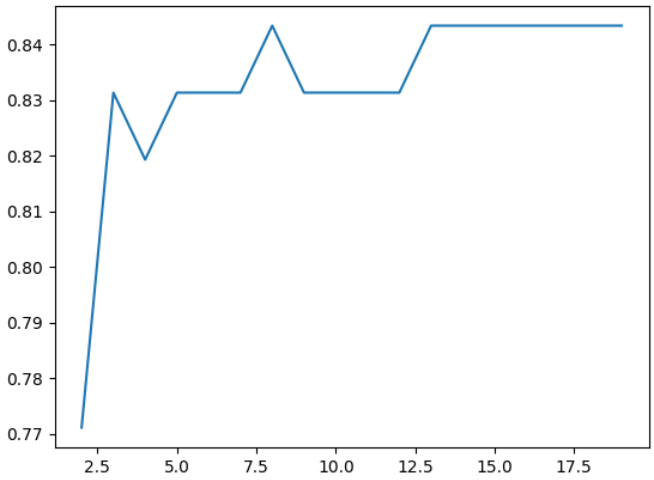

Fig. 17. PCA results with columns (Dengue, Logistic Regression)

Table 1. Accuracy on B-cells of Dengue when PCA = 7

Model	Accuracy
Logistic Regression	79.34%
Decision Tree	88.98%
Random Forest	88.98%
SVM	86.57%
Gradient Boosting	89.39%

Table 2. Accuracy on B-cells of Dengue without PCA

Model	Accuracy
Logistic Regression	78.13%
Decision Tree	82.95%
Random Forest	85.98%
SVM	85.36%
Gradient Boosting	87.77%

6. CONCLUSION

Our work concludes that there is enough physiochemical information in aggregate-based features of an epitope, that it can be differentiated from a completely random epitope generated using learning algorithms. Although the above can simply be because the random samples in question had some property in un-natural bounds. We also conclude that the features calculated like this pose a high level of redundancy when analyzed. Classification accuracy (followed by PCA) does not go downhill even after halving the number of features.

7. FURTHER WORK

Some of the future work to be done is to use the features of the base protein and, hence incorporate the structural information, which has yielded good results, according to the review of past literature.

We will also incorporate thermodynamic stability, among other features, to be better understand the extent of our current work.

We have written a scraper for extracting the calculated features for a given protein sequence and we will use it along with the extracted epitope features for doing epitope prediction for a given protein sequence. The current method we're discussing for implementing the same is a sequential deep learning model which makes prediction over each residue on whether or not the residue will constitute the epitope.