# Reverse Vaccine Design for Anthrax Toxin

## Project by Muhammad Faizan Khalid

### Introduction:

Anthrax is one of the deadliest bioweapons in the world. Anthrax is the colloquial name for the bacterium *Bacillus anthracis*, which can be weaponized as spores that can be inhaled. Its main mechanism of action is the toxin the bacterium produces which can be lethal on its own to eukaryotic cells. The Anthrax toxin is comprised of three non-toxic proteins that combine on eukaryotic host cell surfaces to make noncovalent, toxic complexes1. These proteins include Lethal Factor (LF), Edema Factor (EF) and Protective Antigen (PA).

As part of a reverse vaccine engineering workflow, this document will go through and highlight the key steps and outputs for the design of a hypothetical toxoid vaccine targeting the PA component of the Anthrax toxin. The goal of the vaccine would be to design an antigen based of the PA protein in the toxin, to illicit an immune response in the host so that any future exposure would result in the Anthrax toxin being neutralized by the bodies natural immune system. PA was selected as the target for a vaccine due to its low lethality on its own and functionality in cell binding which results in subsequent self-assembly into a ring-shaped homo-oligomer pre-channel that allows entry of the LF and EF components of the toxin into the cell cytosol. Leading to death of the host cell1.

### Data Retrieval and Sequence Availability:

The first step in reverse vaccine development is obtaining candidate sequences for the target antigen molecule through either genomics analysis or via looking through public databases. In the case of the Anthrax PA protein, the sequence for the protein is well defined and available online through the NCBI protein database2. In this workflow, three different candidate sequences are used, with one sequence derived from a metagenomics assembly3. This was done to contrast structures of the available antigenic proteins and compare between reference and wild type proteins for the purposes of developing a theoretical vaccine.

#### Protein 1:

>WP\_338259392.1 anthrax toxin protective antigen, partial [Bacillus anthracis]

MKKRKVLIPLMALSTILVSSTGNLEVIQAEVKQENRLLNESESSSQGLLGYYFSDLNFQAPMVVTSSTTG

DLSIPSSELENIPSENQYFQSAIWSGFIKVKKSDEYTFATSADNHVTMWVDDQEVINKASNSNKIRLEKG

RLYQIKIQYQRENPTEKGLDFKLYWTDSQNKKEVISSDNLQLPELKQKSSNSRKKRSTSAGPTVPDRDND

GIPDSLEVEGYTVDVKNKRTFLSPWISNIHEKKGLTKYKSSPEKWSTASDPYSDFEKVTGRIDKNVSPEA

RHPLVAAYPIVHVDMENIILSKNEDQSTQNTDSQTRTISKNTSTSRTHTSEVHGNAEVHASFFDIGGSVS

AGFSNSNSSTVAIDHSLSLAGERTWAETMGLNTADTARLNANIRYVNTGTAPIYNVLPTTSLVLGKNQTL

ATIKAKENQLSQILAPNNYYPSKNLAPIALNAQDDFSSTPITMNYNQFLELEKTKQLRLDTDQVYGNIAT

YNFENGRVRVDTGSNWSEVLPQIQETTARIIFNGKDLNLVERRIAAVNPSDPLETTKPDMTLKEALKIAF

GFNEPNGNLQYQGKDITEFDFNFDQQTSQNIKNQLAELNVTNIYTVLDKIKLNAKMNILIRDKRFHYDRN

NIAVGADESVVKEAHREVINSSTEGLLLN

#### Protein 2:

>WP\_154166101.1 anthrax toxin protective antigen, partial [Bacillus anthracis]

MKKRKVLIPLMALSTILVSSTGNLEVIQAEVKQENRLLNESESSSQGLLGYYFSDLNFQAPMVVTSSTTG

DLSIPSSELENIPSENQYFQSAIWSGFIKVKKSDEYTFATSADNHVTMWVDDQEVINKASNSNKIRLEKG

RLYQIKIQYQRENPTEKGLDFKLYWTDSQNKKEVISSDNLQLPELKQKSSNSRKKRSTSAGPTVPDRDND

GIPDSLEVEGYTVDVKNKRTFLSPWISNIHEKKGLTKYKSSPEKWSTASDPYSDFEKVTGRIDKNVSPEA

RHPLVAAYPIVHVDMENIILSKNEDQSTQNTDSQTRTISKNTSTSRTHTSEVHGNAEVHASFFDIGGSVS

AGFSNSNSSTVAIDHSLSLAGERTWAETMGLNTADTARLNANIRYVNTGTAPIYNVLPTTSLVLGKNQTL

ATIKAKENQLSQILAPNNYYPSKNLAPIALNAQDDFSSTPITMNYNQFLELEKTKQLRLDTDQVYGNIAT

YNFENGRVRVDTGSNWSEVLPQIQETTARIIFNGKDLNLVERRIAAVNPSDPLETTKPDMTLKEALKIAF

GFNEPNGNLQYQGKDITEFDFNFDQQTSQNIKNQLAELNATNIYTVLDKSKLNAKMNILIRDKRFHYDRK

NIAVGAEESVVKEAHREVINTSTEGLLLNIDKDIRKILSGYIVEIEDTEGLK

#### Protein 3:

>PNS63514.1 anthrax toxin protective antigen [Bacillus anthracis]

MKKRKVLIPLMALSTILVSSTGNLEVIQAEVKQENRLLNESESSSQGLLGYYFSDLNFQAPMVVTSSTTG

DLSIPSSELENIPSENQYFQSAIWSGFIKVKKSDEYTFATSADNHVTMWVDDQEVINKASNSNKIRLEKG

RLYQIKIQYQRENPTEKGLDFKLYWTDSQNKKEVISSDNLQLPELKQKSSNSRKKRSTSAGPTVPDRDND

GIPDSLEVEGYTVDVKNKRTFLSPWISNIHEKKGLTKYKSSPEKWSTASDPYSDFEKVTGRIDKNVSPEA

RHPLVAAYPIVHVDMENIILSKNEDQSTQNTDSQTRTISKNTSTSRTHTSEVHGNAEVHASFFDIGGSVS

AGFSNSNSSTVAIDHSLSLAGERTWAETMGLNTADTARLNANIRYVNTGTAPIYNVLPTTSLVLGKNQTL

ATIKAKENQLSQILAPNNYYPSKNLAPIALNAQDDFSSTPITMNYNQFLELEKTKQLRLDTDQVYGNIAT

YNFENGRVRVDTGSNWSEVLPQIQETTARIIFNGKDLNLVERRIAAVNPSDPLETTKPDMTLKEALKIAF

GFNEPNGNLQYQGKDITEFDFNFDQQTSQNIKNQLAELNATNIYTVLDKIKLNAKMNILIRDKRFHYDRN

NIAVGADESVVKEAHREVINSSTEGLLLNIDKDIRKILSGYIVEIEDTEGLKEVINDRYDMLNISSLRQD

GKTFIDFKKYNDKLPLYISNPNYKVNVYAVTKENTIINPSENGDTSTNGIKKILIFSKKGYEIG

These will be the sequences for the target antigenic protein and used in subsequent functional analysis. Typically, data mining can be used to study and contrast hundreds of sequences for a candidate antigen simultaneously. Three antigens is an adequate number to show variety and selection for a sample workflow.

### Functional Analysis of Antigenic Proteins:

#### Prediction of antigenicity:

In vaccine design, high antigenicity is important for target recognition. Antigenicity refers to a substance’s ability to be recognized by the immune system and its components like B-cell receptors, antibodies and T-cell receptors. Put simply, it is a measure of how well a protein or molecule functions as an antigen.

For this workflow, the antigenicity of the candidate sequences is tested using three different online tools VaxiJen v.24,5, VaxiJen v.36-8 and the antigenic.pl tool made by the Immunomedicine Group9. These are alignment free methods of predicting antigenicity and rely on machine learning or algorithmic methods to determine immunogenicity for a candidate sequence4-9. Each tool was run with its default settings, and the sequence was set to be bacterial in origin. The results for each sequence are presented in **Table 1**.



Table 1: Table of antigenicity predictions for candidate proteins using three different tools. Protein 1 has higher scores across predictions but protein 3 has the most determinants.

From the results in **Table 1**, it can be observed that all three of the selected candidates are probable candidates for vaccine antigens. Notably, Protein 1 has the higher score across metrics but protein 3, which is the wild type, has more antigenic determinants. Figures for antigenic scores by protein sequence numbers are presented in the appendix (*Appendix* *Figures 1-3*) and data tables from the outputs are included in supplementary materials.

#### Prediction of allergenic response to antigen (Allergenicity):

Allergenicity refers to the ability of a protein or molecule to trigger the body’s immune system to produce IgE antibodies and cause hypersensitivity reactions, such as rashes, asthma, or even anaphylaxis. In simple terms it is a measure of an antigens ability to cause an allergic reaction. Ideally, a vaccine candidate should not elicit any allergic response from the host and should not be considered an allergen. The immune response should be directed at producing targeted antibodies and recruiting the right kind of T and B-cells to remove the pathogen from the host.

To predict allergenicity, several tools exist and it’s best to use more than one to ensure that the vaccine does not cause significant harm to the recipient. In this workflow AlgPred10 was one of three tools used, and it flagged all of the candidates as allergens. This would typically cause concern and requires deeper analysis which could disqualify the antigen candidates for a vaccine. However, since this is a theoretical workflow with only three candidates, and the other tools used did not corroborate AlgPreds10assessment, its results can be ignored, and the workflow can continue.

Notably, all three candidates were not considered allergens by AllerCatPro 2.0 11 and AllerTOP v. 2 12. Two out of three methods used to predict allergenicity did not consider the candidate PA protein sequences as allergens. Scores for AlgPred10 (using SVM module based on amino acid composition) predictions are provided in the appendix (*Appendix* *Table 1*). The scores for the other two methods are included in **Table 2**.

Table 2: Table of allergenicity predictions for candidate proteins using two different tools. All three candidates pass and are marked as non-allergens.



Two of three methods utilized in this analysis indicate that the candidate proteins are non-allergens. While their peptide sequences can later be evaluated for toxicity on their own and modified during manufacturing, this is enough evidence to proceed with all three of the vaccine candidates to the next step in the workflow.

#### Analysis of physio-chemical properties of antigen proteins:

Physio-chemical proteins are an important factor to consider in antigen design. The hydrophobicity, half-life and composition of amino acids play a key role in how stable a molecule is. A vaccine candidate should be stable, have a reasonable half-life and not have strong charges which can hinder its binding to immune cells.

It is important to note that like much of the analysis up till this point in the workflow, the iso-electric point predictions at this stage are theoretical because they are based on the sequences of amino acids, which includes residues that may be inaccessible when the protein is folded into its 3D structure. The folding of the protein will impact which residues are facing the surface of the protein and change the final properties the protein molecule will practically display. Further analysis will be required for the 3D structures of the antigen proteins.

All three candidate antigens were analyzed using ProtParam13 and AA-Prop14 with the results indicating that they form stable products with half-lives of around 30hrs according to predictions. This means they are good candidates for the vaccine design. Additionally, the tools have provided key metrics such as light absorption, extinction rates and molecular weight which can later be used when purifying and isolating the products during the vaccine manufacturing process. The key metrics from this analysis are presented in **Table 3** and additional metrics and hydrophobicity plots are included in supplementary materials**.**



Table 3: Summary table of key physio-chemical properties of candidate proteins. All three are stable molecules with good half-lives.

### Analysis of Secondary and Tertiary Structures of Proteins:

Protein structures have levels, and the chain of amino acids is only the primary structure. The candidate antigens need to be evaluated based on their secondary and tertiary structures. The conformation of the proteins and their folding is what imparts functionality to the final 3D structure, and this is important in determining if an antigen will have meaningful interactions with the immune system when used as a vaccine.

Key metrics at this stage of the design process are the number of α-helixes, β-strands, disordered regions, functional domains and solvent access. We already know from the previous analysis that the three proteins are stable. However, the selection of antigens can be narrowed down at this stage based off the analysis of secondary and tertiary structures using Phyre215, and the prediction of folding and disordered regions using PSIPRED16-20. The summary of these results is presented in **Table 4**.

Table 4: Summary from analysis of secondary and tertiary structures of candidate antigen proteins. Protein 1 has the highest number of predicted domains combined with the lowest predicted percentage of disordered regions by Phyre215.

Using alignments to pfam via the InterPro21 server, the sequences can be mapped to references to identify the functional domains present in the antigen candidates. It is known from literature that the anthrax PA protein has 4 domains that have been experimentally studied and observed1. Only 2 of the candidate molecules were found to have coverage across 2 domains via InterPro21, while all 3 showed the structural presence of all 4 via the Phyre215 predicted models.

The 4 known domains include **Domain 1**: Activation domain (furin site, Ca²⁺ binding), **Domain 2**: Pore-forming/oligomerization, **Domain 3**: Connector (structural hinge) and **Domain 4**: Receptor-binding (Ig-like) domain. The Ig-like domain makes, **Domain 4** a desirable region for a vaccine candidate and only protein 1 was confirmed by pfam alignment to have that domain in its sequence. Ultimately, while all three proteins are predicted to contain the full PA structure, Protein 1 is the most comprehensively annotated, has lower disorder, and clearer domain resolution, particularly for **Domain 4**, the critical receptor-binding domain. These characteristics and its reduced likelihood of being allergenic and higher antigenicity, make it the most promising candidate for epitope mapping and vaccine design.

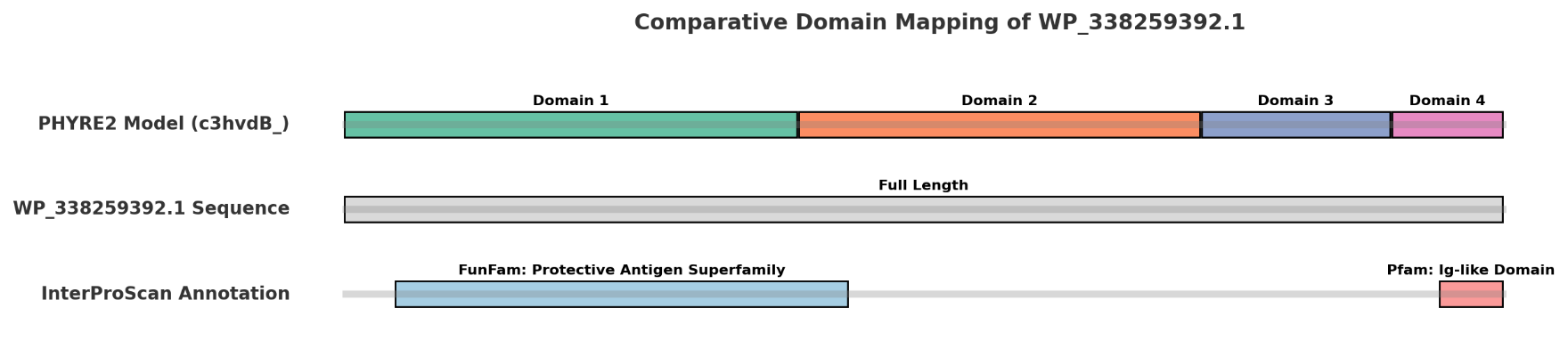
From this analysis we can select Protein 1 as the candidate to move forward in the workflow. All InterPro21 and PSIPRED16-20 results alongside predicted pdb models from the Phyre215 analysis are provided as supplementary materials and a summary of domain structures and InterPro21 mapping to pfam in protein 1 is presented as a cartoon in **Figure 1.**

Figure 1: Summary cartoon showing domain structures and mapping for anthrax PA protein, Protein 1, structure prediction via Phyre215 and InterPro Annotation21. Protein 1 (WP\_338259392.1) has the key domain desirable for a vaccine candidate, the Ig-like domain and structurally contains the other domains based of the predicted structure even if it may not be annotated by alignment.

### Analysis and Prediction of Epitope-Antigen Interactions:

The next step in the vaccine design process is to model and predict the epitopes on the candidate antigen. Epitopes are the key binding regions on the antigen molecule that will interact with the host immune system. Peptide sequences in the case of proteins. There are several tools to do this as there are multiple cell types to consider with their own unique epitopes on the target antigen. From the previous analysis we know Protein 1 is an ideal candidate with good stability and antigenic properties, so it can be selected for epitope prediction in the vaccine design process.

For a toxoid vaccine like the one in consideration here which is derived from the toxin of a bacteria, the key epitopes to be modeled are B-cell epitopes for continuous (amino acid chain) and conformational or discontinuous (3D structure) forms of the antigen. With B-cells being responsible for producing the antibodies against the toxin. Secondly, there are T-cell epitopes which are important to predict for the hosts long term specific immune responses. T-cell epitope modeling also involves modeling MHC class 2 epitopes in the process as the antigen will be engulfed by Antigen Presenting Cells (APCs) that will process the antigen, signal the T-cell and present the antigens to the CD4+ T-cells.

#### Prediction of continuous or sequential B-cell epitopes:

To predict the continuous or sequential B-cell epitopes, the sequence of Protein 1 was submitted to two different servers for prediction and scoring, and the aggregate of both predictions are presented in **Figure 2**. For the prediction of epitopes, the two methods used were The Immune Epitope Database (IEDB)22 and BepiPred-2.023. Both methods have similar outputs with BepiPred-2.023 providing additional metrics and information about the residues in the protein structure.

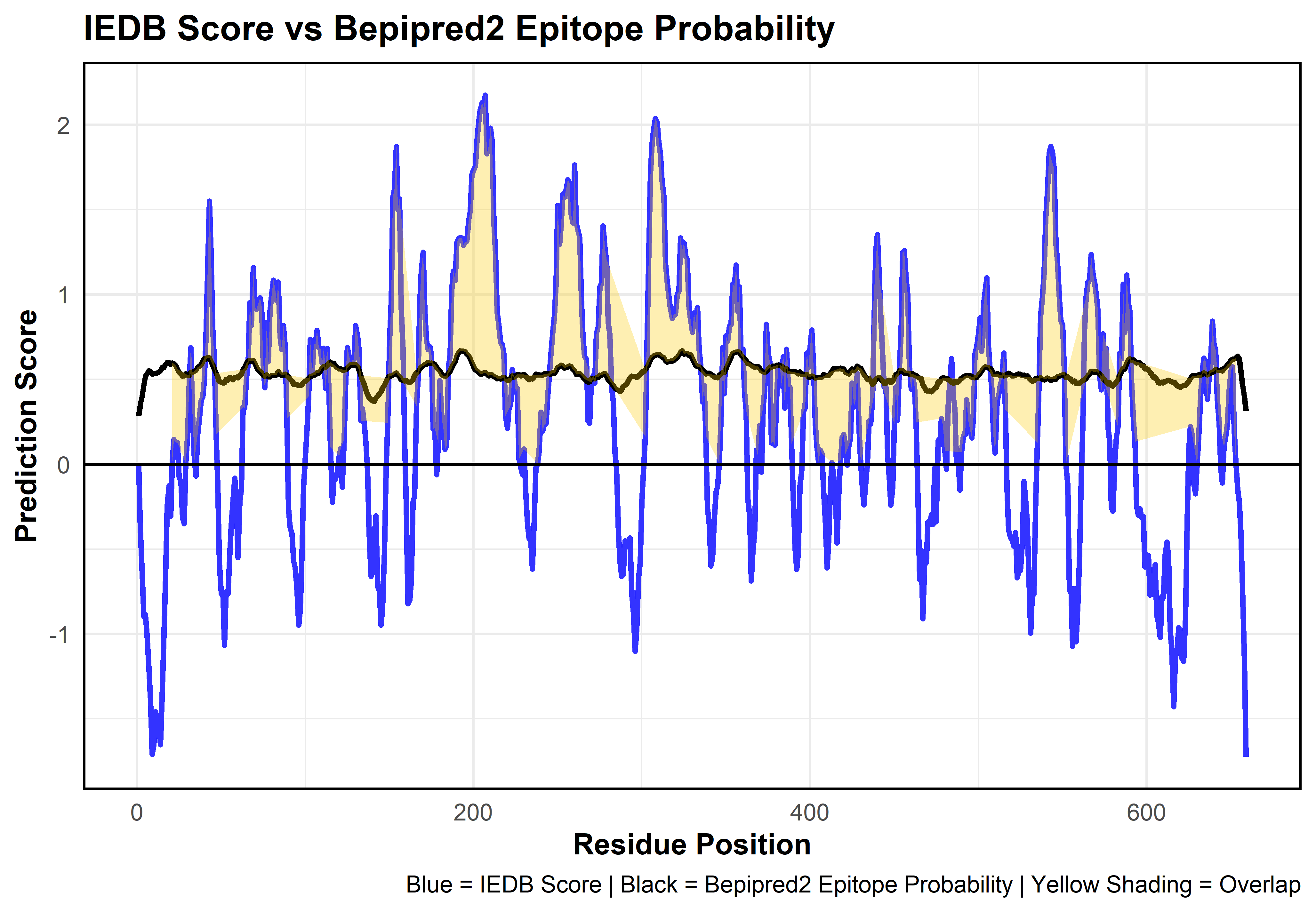


Figure 2: Figure showing combined scores from IEDB22 and BepiPred-2.023 overlayed on the same figure. The yellow shaded region indicates, areas of high overlap above thresholds (IEDB Score > 0 and BepiPred-2.0 Epitope Probability > 0.5). These are the sections of the linear antigen most likely to have epitopes. Regions bellow 0 are unlikely to have any epitopes.

Both methods showed high concordance and indicated the presence of epitopes in the linear structure of the antigen. The analysis of the results from both tools and the visual analysis was conducted in R and all scripts from this analysis are available in the supplementary materials. Additionally, the epitope prediction scores from both methods were averaged to identify the top five epitopes, their sequences were preserved for downstream analysis, and they were mapped to the four predefined domains of the PA protein to see where the key binding incidents would take place based on the predictions and domain boundaries. The results from this analysis are presented in **Table 5**.

Table 5: Table showing averaged scores, sequences and mapped domains for the top 5 predicted epitopes from the analysis performed using IEDB22 and BepoPred-2.023. Epitopes are also ranked by their top average score across both tools.

#### Prediction of conformational or discontinuous B-cell epitopes:

To determine conformational epitopes, it is important to first have an accurate 3D model for the structure of an antigen protein. In previous steps, Phyre215 provided a 3D model, and that can be used for this analysis. However, to confirm that model’s validity, SWISS-MODEL24 was also used to generate a 3D structure for Protein 1 using its sequence. Notably, the model generated by SWISS-MODEL was different from the one predicted by Phyre2 (Phyre 215: c3hvdB 99% identity/100% Confidence, SWISS-Model24: 1t6b 100% identity/ 99.85% Coverage). We need an accurate 3D representation to be able to identify the outer structures of a protein and then select the peptides on the outer regions which are most likely to bind to immune cells.

To resolve this conflict and choose the best 3D structure, the models were compared via BLASTp25,26 alignment to the PDB database. The top hit was PDB: 1T6B\_X with 100% identity, indicating that the SWISS-MODEL was more accurate because it matched accurately and completely to the full sequence in the reference database. The results from BLASTp25,26 and PDB files for both models are included in supplementary materials.

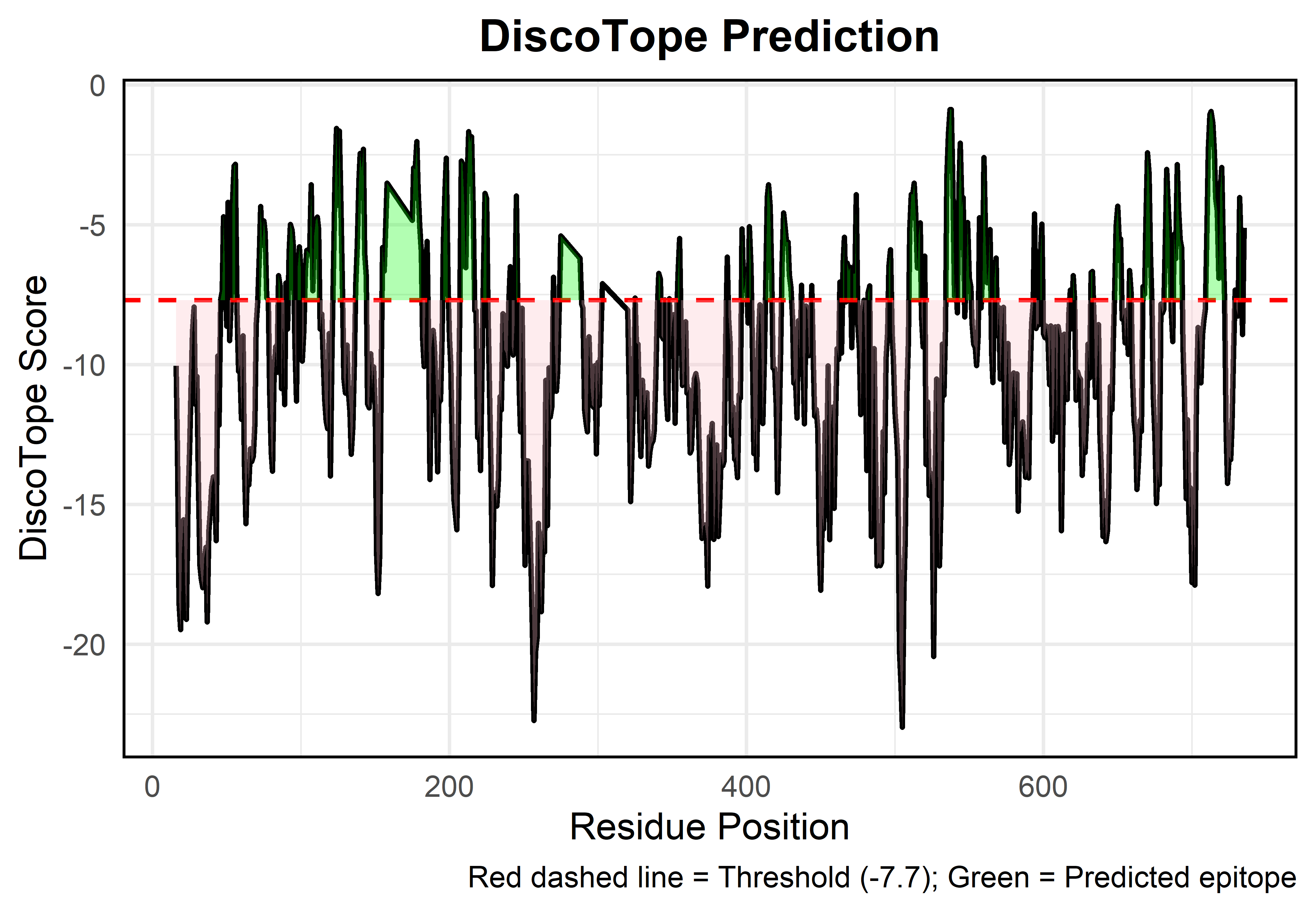
To predict epitopes on the 3D structure, DiscoTope27 hosted on the IEDB server was utilized. The predictions were plotted by residue in **Figure 3** and the epitopes were filtered using R to identify the top five to match the previous analysis for linear epitopes and is summarized in **Table 6**. The scripts used to generate the figures and analysis along with the raw data output from DiscoTope27 are included in supplementary materials.

Figure 3: Figure showing DiscoTope27 Score by Residue. The green shaded portions indicate the presence of conformational epitopes based on DiscoTope27 score threshold of -7.7. There are several predicted epitopes.

Table 6: Table showing average score, sequences and mapped domains for the top five predicted epitopes from the analysis performed using DiscoTope27. Epitopes are ranked based on score.



There are ample conformational epitopes in the candidate protein and notably, the top five epitopes include an epitope that maps to **Domain 4** which is known from structural analysis to be a binding domain for an antibody like peptide. A promising result for a vaccine candidate.

#### Prediction of T-cell and MHC Class II peptide binding:

When considering binding and immune responses for T-cells as in the case of a toxoid vaccine, it is important to keep in mind that CD4+ T-cells interact with MHC class II complexes on APCs. To model T-cell immune responses it is first necessary to identify peptide interactions between the antigen and MHC class II complexes. Class I complexes are not relevant in the case of a toxoid vaccine because they are responsible for calling CD8+ killer T-cells to infected or cancerous cells to trigger cell death. The toxin antigen is more likely to be engulfed by an APC, processed and then complexed with MHC class II molecules for generating a specific immune response.

Since the MHC class II molecule is expressed by the human genome, it has various alleles that need to be considered. Each allelic variation has its own set of interactions with the antigenic peptides that need to be predicted. To predict the peptide binding and identify promising T-cell epitopes for inclusion in a toxoid vaccine, comprehensive MHC class II binding predictions were made using two independent and widely used platforms IEDB MHC-II Binding Predictions (NetMHCIIpan 4.1)28 and TepiTool29. Both tools were configured to evaluate 15-mer peptides derived from the sequence of Protein 1 across a broad panel of 25 human HLA-DR, -DP, and -DQ alleles, ensuring population-wide relevance. A prediction threshold of ≤10% percentile rank was applied to identify strong binders, and filtering was done on the results to find a consensus set of peptides across both tools.

The consensus epitope list generated by cross-referencing peptides that were classified as strong binders in both platforms is provided in the supplementary materials along with analysis scripts, raw and summary outputs from each tool. Each shared peptide was scored using a consensus score that was the calculated average of its predicted binding ranks from IEDB28 and TepiTool29. Lower consensus scores indicate higher predicted binding affinity and confidence. Peptides that ranked highly across both tools were prioritized for further consideration and the top five peptides are presented in **Table 7** and as a heat map in **Figure 4**.



Table 7: Table of top five consensus peptides predicted to have high binding affinity to MHC Class II complexes by IEDB28 and TepiTool29. Lower consensus score indicates better binding outcomes in vaccine design. Peptides are sorted by rank from top to bottom based on consensus score and IC50 values. The best is at the top of the list.

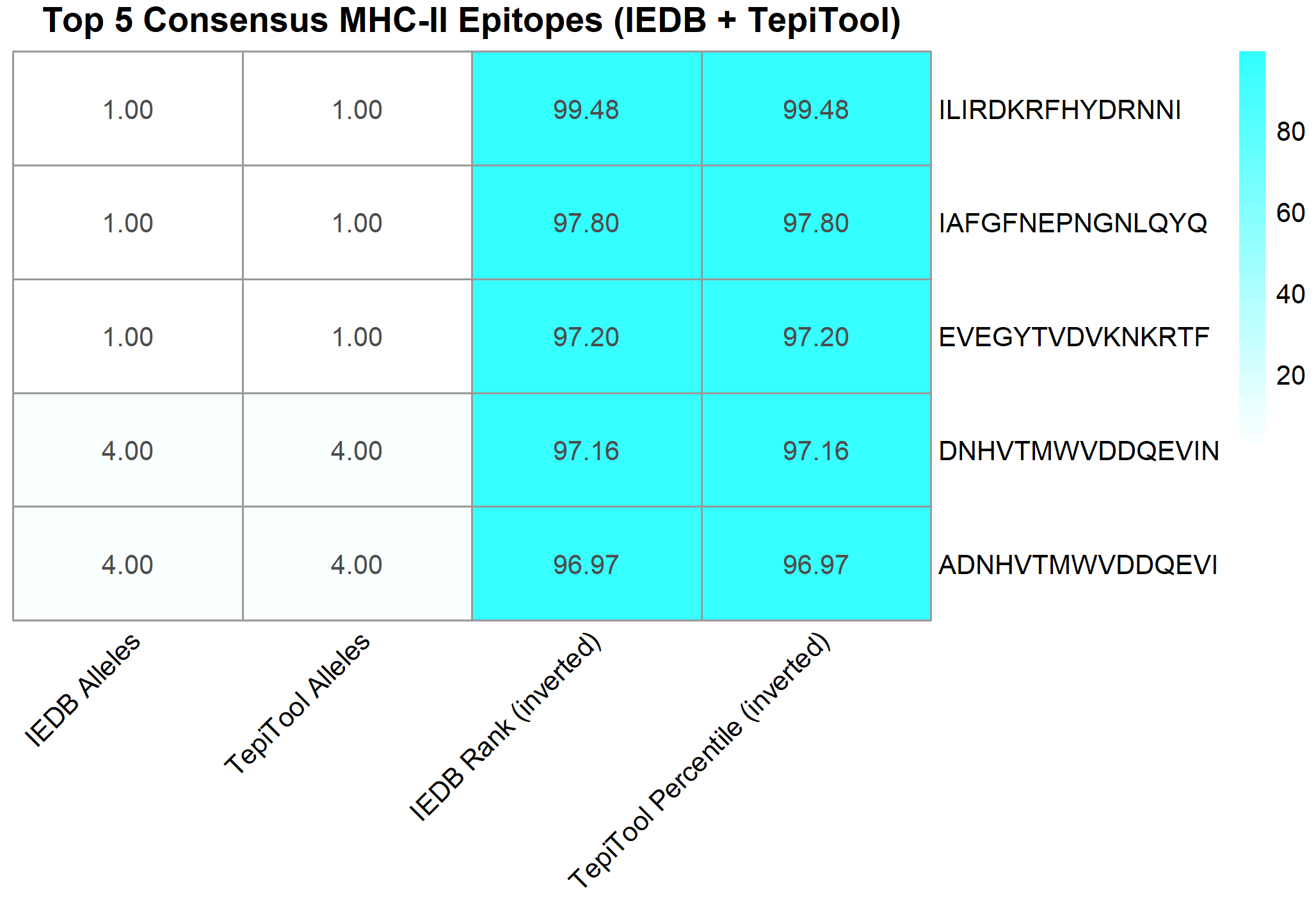


Figure 4: Heatmap of top five consensus MHC class II peptides based on IEDB28 and TepiTool29 predictions. Binding scores from IEDB (rank) and TepiTool (percentile) were inverted to percentage binding strength, with higher values indicating stronger predicted affinity. The number of HLA alleles each peptide binds to in both tools is also shown.

The presence of highly promiscuous peptides that bind multiple HLA alleles across platform predictions with low IC50 and percentile rank scores, indicating strong binding predictions, show promising results for Protein 1 being a potential vaccine candidate. Successful binding to MHC Class II molecules also gives us a good indication of T-cell immune response as the receptors on T-cells will most likely respond to antigen peptides presented by the MHC class II molecules if presented by APCs.

#### Prediction of CD4+ T-cell immunogenicity:

Using IEDBs CD4 T-cell immunogenicity prediction tool30, it is possible to identify the immunogenicity of our antigen for at least seven of the most common HLA alleles in the world. Using seven alleles, the tool is able to create an aggregate immunogenicity score for 15mer peptide sequences in the antigen. Thus, it is possible to confirm that CD4 cells will interact with the target antigen if presented by an MHC II complex or potentially even in the humoral immune response. All identified peptides and their scores are included in supplementary materials, but the top five peptides identified in the antigen and their immunogenicity scores are presented in **Table 8.**



Table 8: Predicted immunogenicity for top five peptides in Protein 1 across seven most common alleles for HLA genes relevant for CD4+ T-cell immune responses. Higher scores indicate a stronger immune response from CD4+ T-cells. Median percentile rank is across the seven most common HLA genes which include: HLA-DRB1:03:01, HLA-DRB1:07:01, HLA-DRB1:15:01, HLA-DRB3:01:01, HLA-DRB3:02:02, HLA-DRB4:01:01, and HLA-DRB5:01:01.

The high CD4 + T-Cell immunogenicity scores are promising and validate our analysis thus far that Protein 1 can function as a good vaccine candidate. With ample evidence that the candidate vaccine antigen will illicit an immune response and most likely have meaningful interactions with almost every category of immune cell necessary in a toxoid vaccine response mechanism, the workflow can proceed to modeling the epitopes on the protein structure and visual analysis.

#### Modeling of predicted epitopes:

Once a comprehensive analysis of epitopes has been conducted for the candidate antigen. The next step is to model the predicted epitopes on to the surface of the 3D structure of the molecule. This will allow visualization of where the antibodies will bind to the protein in the vaccine antigen. From the previous steps, a highly reliable 3D structure was already generated for the antigen using SWISS-MODEL24. In this case, the antigen protein was determined to have very high similarity to the PDB protein 1t6b. Using this information, the tool ElliPro31 can predict where in the molecule antibodies are most likely to bind based off the antigens structure.

In previous steps DiscoTope27 was used to predict discontinuous/conformational epitopes with Epitope Database (IEDB)22 and BepiPred-2.023 being used to predict continuous epitopes. ElliPro31 offers some redundancy and complementarity in that it will recalculate both. However, it offers new information and advantages in its methods and outputs compared to previous tools that provide per-residue scores rather than clearly defined epitope regions. Overall, ElliPro31 offers other key advantages that make it better suited for structural modeling and visualization of discrete antibody-binding regions or epitopes.

Most notably, Ellipro31 clusters residues into defined epitope regions (continuous and discontinuous), directly identifying likely immunogenic patches using an ellipsoid approximation algorithm to define the protein's 3D envelope and assigns a protrusion index to each residue. This provides a more holistic view of surface-exposed and protruding epitopes. Ultimately, giving a better understanding of regions accessible to solvents and functionally useable as epitopes. Lastly, its output can be easily visualized with tools such as UCSF ChimeraX32. All outputs from ElliPro31 and scripts including cxc commands used to generate the model figures are presented in supplementary materials. Continuous epitopes on the surface of Protein 1 are presented in **Figure 5** and discontinuous epitopes are presented in **Figure 6.** Additional figures with the molecular backbone removed and only the epitopes highlighted are included in the appendix to allow easy visualization of just the epitopes (*Appendix Figure 4, 5*).

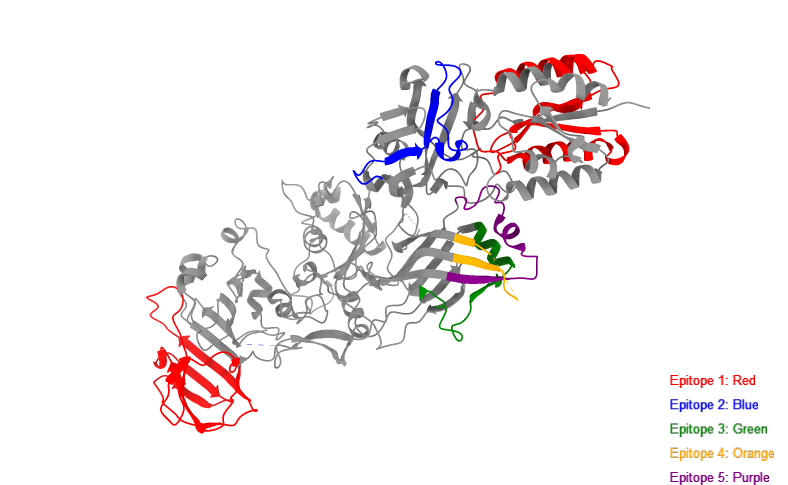


Figure 5: 3D structure of Protein 1 with continuous epitopes predicted by ElliPro31 highlighted in ranked order. Ranking is based on protrusion index (PI) score which indicates how accessible and immunogenic the epitope is. The top five epitopes are shown with Epitope 1 being the most immunogenic and likely to bind an antibody (score = 0.8) and Epitope 5 being the least likely (score = 0.7). There were 16 total continuous epitopes predicted.

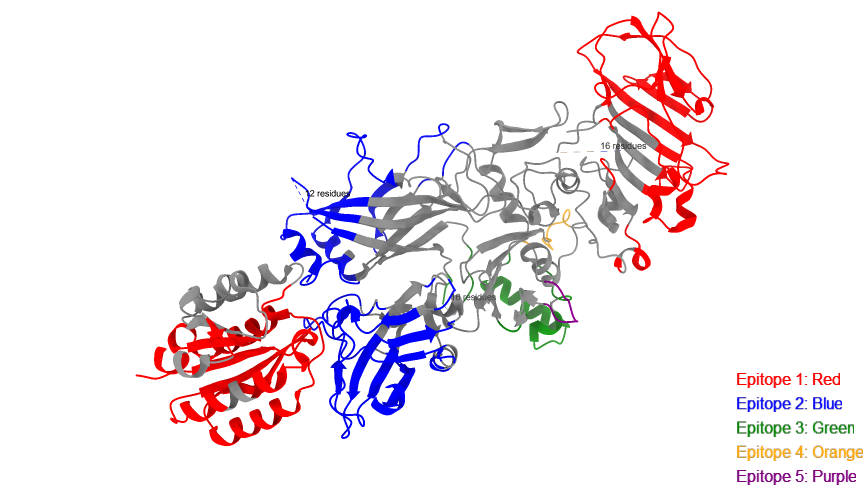


Figure 6: 3D structure of Protein 1 with discontinuous epitopes predicted by ElliPro31 highlighted in ranked order. Ranking is based on protrusion index (PI) score which indicates how accessible and immunogenic the epitope is. The top five epitopes are shown with Epitope 1 being the most immunogenic and likely to bind an antibody (score = 0.74) and Epitope 5 being the least likely (score = 0.5). Only 5 discontinuous epitopes were predicted in total.

With the modeling completed, the epitopes on the antigen candidate can be highlighted and antibody binding can be visually confirmed and observed on the molecule to illicit an immune response. High scoring epitopes (Score between 0-1) predicted by ElliPro31, are a positive indicator of immunogenicity. The high score means that the epitope is outward from the protein’s 3D envelope and a likely target for antibodies. There are an ample number of high scoring epitopes in the model, and this is corroborated by previous analysis. The model structure can be further worked on to visualize molecular docking with antibody molecules and the toxicity profile of the highlighted epitopes can also be evaluated for final vaccine production to filter out any toxic regions or manufacture additional epitopes to be inserted into the protein sequence as needed.

For the purposes of this workflow, however, it is feasible to proceed to the next stage which is population coverage analysis. Manufacturing considerations, further toxicity analysis and protein modifications are key steps in drug design but exceed the scope of this workflow that focuses on validating an antigen candidate.

### Analysis of Population Coverage of Peptide Relevant Alleles:

In one of the previous analyses when predicting MHC Class II epitopes, the subject of alleles was touched upon. When it comes to global populations, alleles are distributed unevenly across the world and hence immune responses tend to be regulated by the genetic makeup of any given population. During the analysis of MHC II complexes using TepiTool29, part of the output included peptide sequences from the antigen being identified as epitopes and their mapping to known allele variants of MHC class II alleles. This information can be used with the IEDB Population Coverage Calculation33 tool to determine if the epitopes in the candidate antigen will have recognition or coverage in a population. This is a way to determine if the immune response will be elicited equally in different populations. A key step in vaccine design

In this workflow, Pakistan and Canada were chosen as the candidate populations to be studied for coverage. The scripts used to generate the input for IEDBs Population Coverage tool33 from the raw TepiTool29 output as well as all the outputs from the analysis are included in supplementary materials. The population coverage for Pakistan was 1.18% which is negligible and not worth considering but is included in the appendix (*Appendix Figure 6*). However, the coverage for Canada was promising and can be seen in **Figure 7** and is summarized in **Table 9**.

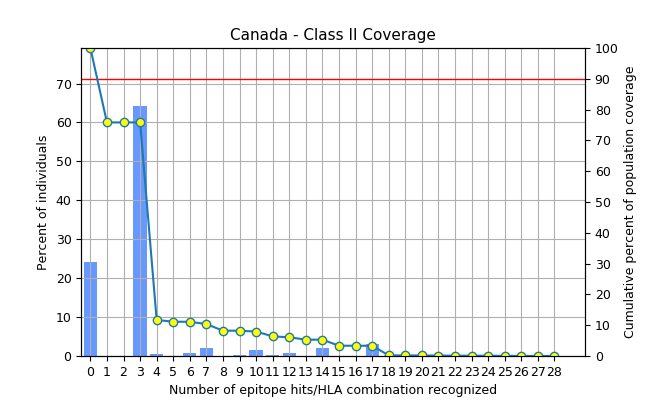


Figure 7: Output from IEDBs population coverage tool33. The results show that Approximately 90% of the Canadian population would likely respond to at least one of the predicted epitopes and the left skew indicates that Individuals commonly recognize 2–4 epitopes based on their HLA gene combinations. This is a positive result for a vaccine candidate.



Table : Summary table of output from IEDBs Population coverage tool33. Good coverage indicates that the candidate antigen is also a good vaccine candidate.



From the high coverage values of around 75.8% It can be concluded that most of the Canadian population will respond to the epitopes in the potential vaccine. Further reinforcing the design and adding validity to Protein 1 and the epitopes identified thus far as a vaccine candidate. While the lack of coverage in the population of Pakistan might indicate the need for a regional modification to the candidate vaccine, resulting in a separate design to illicit an equal immune response in that population with a different set of epitopes. The discrepancy in coverage can result from a lack of relevant alleles in the database being used for epitope prediction. If the database is lacking in alleles relevant to Pakistan, it would also predict lower coverage.

Pakistan is a poorer region of the world where genetic testing isn’t readily available and that can lead to knowledge gaps in allele frequencies amongst the population. To address this, more genetic testing can reveal frequently occurring alleles and exploratory analysis can help identify more relevant alleles to the population. The vaccine may still be effective, but an absence of data can skew the results when it comes to coverage. Notably, when the broader South Asia region was run with the tool the population coverage increased to 98.96% (*Appendix Figure 7*). This rudimentary step shows how including more alleles from the region by selecting a broader dataset can affect the results. India is a much larger and developed country with more information on alleles for its population. It is also a neighbor to Pakistan. Its data can help to elucidate the results and highlight the gap in knowledge for Pakistan and the disparity in genetic testing across populations.

Overall, the results indicate that Protein 1 is rich in epitopes that can illicit meaningful immune responses across diverse populations and has epitopes that are recognizable by a range of immune cells. Making it a viable candidate for reverse vaccine design.

### Conclusion:

Having gone through the steps of designing a toxoid vaccine, it is worth mentioning that vaccines already exist for the bacterium responsible for the toxin34. Before the use of Anthrax as a bioweapon, there were already vaccines available due to agricultural exposure to the bacterium and its spores34. Following the use of the spores as a bioweapon, newer vaccine models were generated that were based on weakened forms of the bacterium and relied on antibiotics to remove the bacteria before it could overwhelm the body with toxins. To date there are several licensed products that can be considered toxin vaccines for anthrax35.

The work presented here is not in any way affiliated with, reflective of or derived from any licensed vaccine product for Anthrax or its toxins. This was done as part of a course project to solidify concepts and learn skills pertinent to immunoinformatic and vaccine design principals. As an initial analysis into a potential vaccine design, this project hopes to highlight key steps in the selection and evaluation of an antigen. The PA protein in Anthrax was chosen due to its benign nature and key function in the ability of the Anthrax toxin to do harm to host cells.

Through the process of reverse vaccine design, candidate antigen sequences were filtered, based on their immunogenicity, allergenicity, structural properties and presence of key immune response regions. The analysis was multifaceted and relied on popular publicly accessible tools that were parse using the R programing language to compare, evaluate and visualize results when required. The candidate proteins were narrowed down to a single high-quality structure which was modeled and evaluated for its immunogenic properties for a vaccine candidate across various populations.

While there are hundreds and thousands of other similar structures to consider as candidates for an Anthrax toxin vaccine, they can be compared against the sequence provided in this workflow as a baseline. The modeled 3D structure can be used to study the docking of antibodies to a potential vaccine. The structure identified can serve as a template for modification and development of a synthetic vaccine as well. Notably, no toxicity analysis was done on the high-quality epitopes identified here. Analysis of those regions for toxic reactions in the body would be necessary to ensure a safety profile for any medical usage.

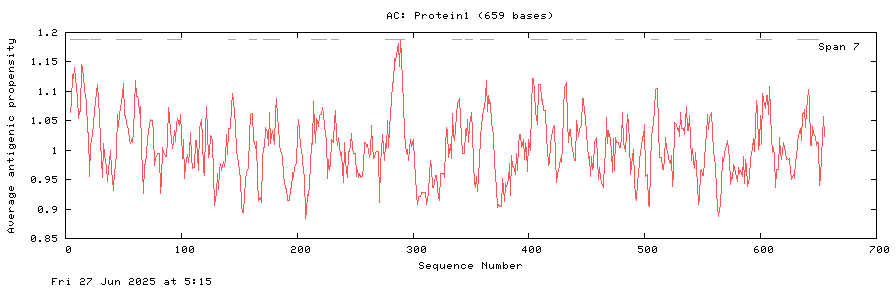
The next logical steps in developing a product would require more detailed modeling and further modifications necessary for manufacturing and synthesis. Steps that exceed the scope of this project. The materials and analysis generated as part of this project however are necessary for any future design aspects and represent a key part of the reverse vaccine design pipeline. Hopefully this analysis reduces barriers in any future vaccine design for Anthrax toxin and is a resource for those looking to learn the process for reverse vaccine design.

### References:

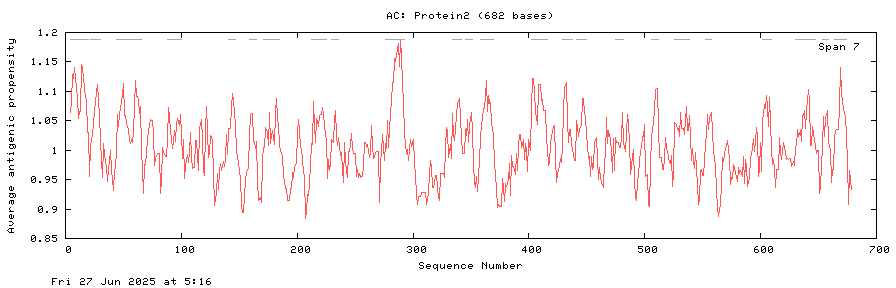
1. Kintzer, A. F., Thoren, K. L., Sterling, H. J., Dong, K. C., Feld, G. K., Tang, I. I., ... & Krantz, B. A. (2009). The protective antigen component of anthrax toxin forms functional octameric complexes. *Journal of molecular biology*, *392*(3), 614-629.
2. Liu, S., Moayeri, M., & Leppla, S. H. (2014). Anthrax lethal and edema toxins in anthrax pathogenesis. *Trends in microbiology*, *22*(6), 317-325.
3. Sahin, M., Buyuk, F., Baillie, L., Wölfel, R., Kotorashvili, A., Rehn, A., ... & Grass, G. (2018). The identification of novel single nucleotide polymorphisms to assist in mapping the spread of Bacillus anthracis across the Southern Caucasus. *Scientific Reports*, *8*(1), 11254.
4. Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*, *8*, 1-7.
5. Doytchinova, I. A., & Flower, D. R. (2007). Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine*, *25*(5), 856-866.
6. Dimitrov, I., Zaharieva, N., & Doytchinova, I. (2020). Bacterial immunogenicity prediction by machine learning methods. *Vaccines*, *8*(4), 709.
7. Doneva, N., & Dimitrov, I. (2024). Viral immunogenicity prediction by machine learning methods. *International Journal of Molecular Sciences*, *25*(5), 2949.
8. Sotirov, S., & Dimitrov, I. (2024). Application of machine learning algorithms for prediction of tumor t-cell immunogens. *Applied Sciences*, *14*(10), 4034.
9. Kolaskar, A. S., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS letters*, *276*(1-2), 172-174.
10. Saha, S., & Raghava, G. P. S. (2006). AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic acids research*, *34*(suppl\_2), W202-W209.
11. Nguyen, M. N., Krutz, N. L., Limviphuvadh, V., Lopata, A. L., Gerberick, G. F., & Maurer-Stroh, S. (2022). AllerCatPro 2.0: a web server for predicting protein allergenicity potential. *Nucleic Acids Research*, *50*(W1), W36-W43.
12. Dimitrov, I., Bangov, I., Flower, D. R., & Doytchinova, I. (2014). AllerTOP v. 2—a server for in silico prediction of allergens. *Journal of molecular modeling*, *20*, 1-6.
13. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. E., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics protocols handbook*, 571-607.
14. AA‑Prop – Protein Physicochemical Properties Prediction Tool. (n.d.). *BioGem*. Retrieved June 30, 2025, from [http://www.biogem.org/tool/aa‑prop/](http://www.biogem.org/tool/aa%E2%80%91prop/)
15. Powell, H. R., Islam, S. A., David, A., & Sternberg, M. J. (2025). Phyre2. 2: A community resource for template-based protein structure prediction. *Journal of Molecular Biology*, 168960.
16. Buchan, D. W., Moffat, L., Lau, A., Kandathil, S. M., & Jones, D. T. (2024). Deep learning for the PSIPRED protein analysis workbench. *Nucleic Acids Research*, *52*(W1), W287-W293.
17. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, *292*(2), 195-202.
18. Lobley, A., Sadowski, M. I., & Jones, D. T. (2009). pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, *25*(14), 1761-1767.
19. McGuffin, L. J., & Jones, D. T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, *19*(7), 874-881.
20. Nugent, T., & Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics*, *10*, 1-11.
21. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., ... & Bateman, A. (2023). InterPro in 2022. *Nucleic acids research*, *51*(D1), D418-D427.
22. Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., ... & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research, 47*(D1), D339–D343. <https://doi.org/10.1093/nar/gky1006>
23. Jespersen, M. C., Peters, B., Nielsen, M., & Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic acids research*, *45*(W1), W24-W29.
24. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., ... & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296-W303.
25. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.
26. Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A., & Yu, Y. K. (2005). Protein database searches using compositionally adjusted substitution matrices. *The FEBS journal*, *272*(20), 5101-5109.
27. Petri, T., Berchtold, E., Zimmer, R., & Friedel, C. C. (2012). Detection and correction of probe-level artefacts on microarrays. *BMC bioinformatics*, *13*, 1-12.
28. Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic acids research*, *48*(W1), W449-W454.
29. Paul, S., Sidney, J., Sette, A., & Peters, B. (2016). TepiTool: a pipeline for computational prediction of T cell epitope candidates. *Current protocols in immunology*, *114*(1), 18-19.
30. Dhanda, S. K., Karosiene, E., Edwards, L., Grifoni, A., Paul, S., Andreatta, M., ... & Sette, A. (2018). Predicting HLA CD4 immunogenicity in human populations. *Frontiers in Immunology*, *9*, 1369.
31. Ponomarenko, J., Bui, H. H., Li, W., Fusseder, N., Bourne, P. E., Sette, A., & Peters, B. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC bioinformatics*, *9*, 1-8.
32. Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science*, *32*(11), e4792.
33. Bui, H. H., Sidney, J., Li, W., Fusseder, N., & Sette, A. (2007). Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC bioinformatics*, *8*, 1-6.
34. Committee to Assess the Safety, & Efficacy of the Anthrax Vaccine. (2002). *The Anthrax Vaccine: Is It Safe? Does It Work?*. National Academies Press.
35. Wang, J. Y., & Roehrl, M. H. (2005). Anthrax vaccine design: strategies to achieve comprehensive protection against spore, bacillus, and toxin. *Medical Immunology*, *4*, 1-8.

### Appendix:

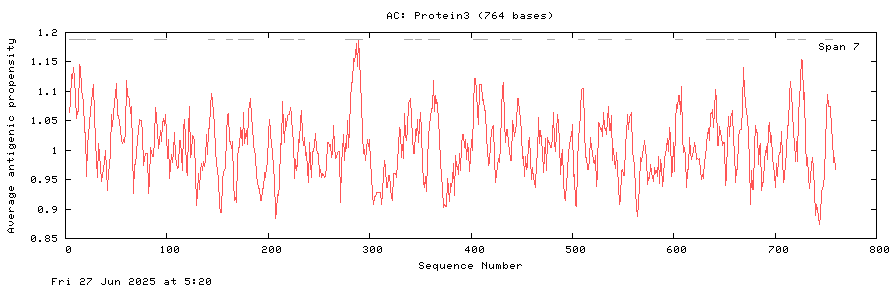
#### Antigenicity per Sequence Graphs:



Appendix Figure 1: Appendix figure for protein 1 showing average antigenicity per sequence.



Appendix Figure 2: Appendix figure for protein 2 showing average antigenicity per sequence.

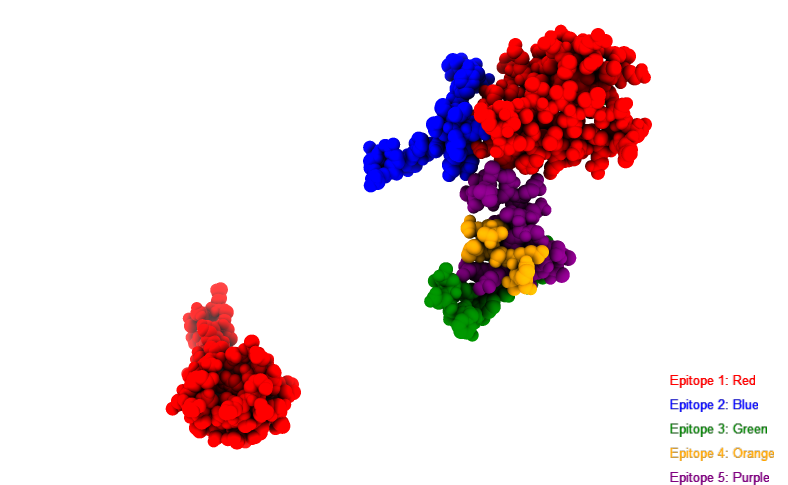


Appendix Figure 3: Appendix figure for protein 3 showing average antigenicity per sequence.

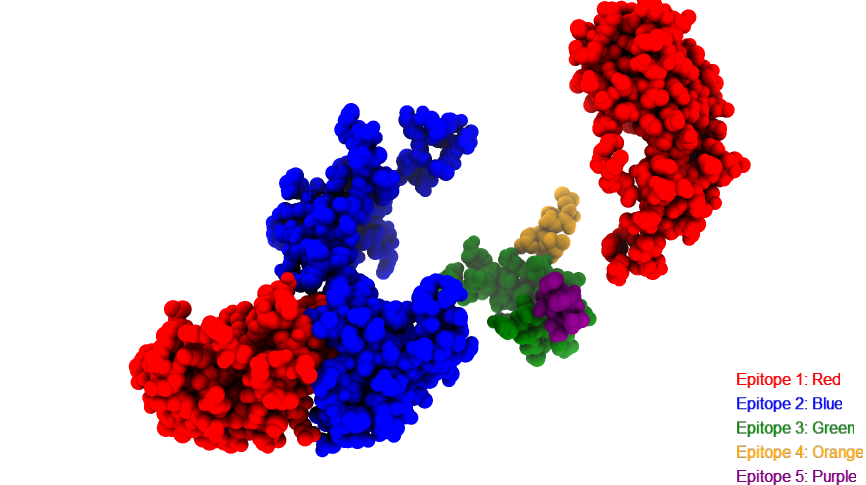
#### Allergenicity Prediction Results:

Appendix Table 1: Table showing results from various allergenicity prediction tools. AlgPred10 is the only method that flags the three candidates as allergens (Threshold 0.4).

#### Epitope Modeling with ElliPro31:

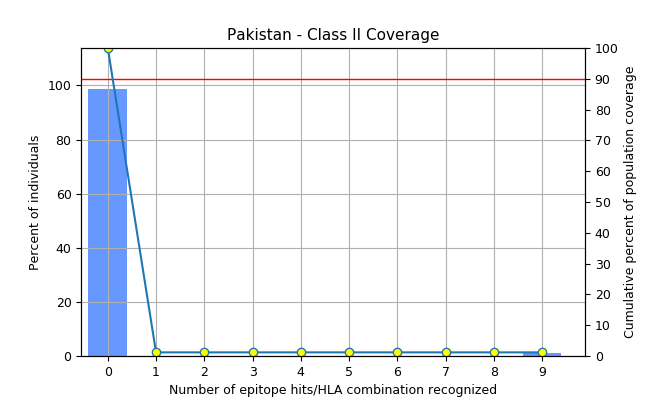


Appendix Figure 4: Appendix figure for 3D structure of Protein 1 with continuous epitopes predicted by ElliPro31 highlighted in ranked order. Main structural backbone of Protein 1 is hidden to only show epitopes. Ranking is based on protrusion index (PI) score which indicates how accessible and immunogenic the epitope is. The top five epitopes are shown with Epitope 1 being the most immunogenic and likely to bind an antibody (score = 0.8) and Epitope 5 being the least in the figure (score = 0.7). There were 16 total epitopes predicted.

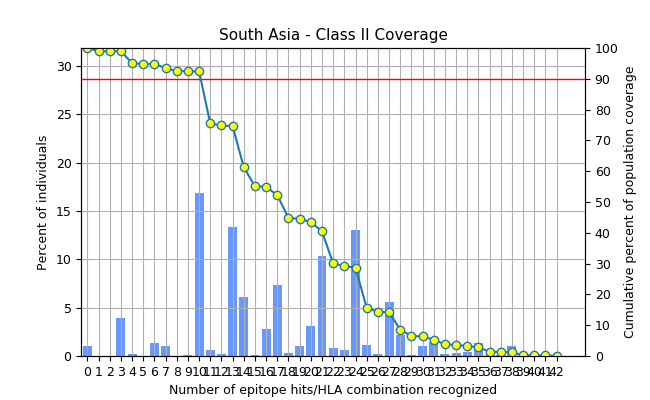


Appendix Figure 5: Appendix figure for 3D structure of Protein 1 with discontinuous epitopes predicted by ElliPro31 highlighted in ranked order. Main structural backbone of Protein 1 is hidden to only show epitopes. Ranking is based on protrusion index (PI) score which indicates how accessible and immunogenic the epitope is. The top five epitopes are shown with Epitope 1 being the most immunogenic and likely to bind an antibody (score = 0.74) and Epitope 5 being the least in the figure (score = 0.5).

#### Population Coverage Figures Pakistan and South Asia:



Appendix Figure 6: Appendix figure from IEDB33 population coverage analysis of the population of Pakistan. The low values of around 0–1%, which are effectively negligible indicate that the selected epitopes are not representative of HLA alleles commonly found in Pakistan. This could also be due to a lack of relevant alleles in the database initially, hence the mapping and subsequent prediction not being possible.



Appendix Figure 7: Appendix figure from IEDB33 population coverage analysis of the population of South Asia as a whole. The values are much higher than for Pakistan with coverage around 98.9%, and the average number of hits is 17.51. This could be evidence of a lack of research on relevant alleles in the database of alleles for Pakistan.