Proposal: [Your project name]

DATA 450 Capstone

[Your Name]

March 1, 2025

1 Introduction

The immune system plays a critical role in identifying and eliminating cancer cells, with T-cell epitopes serving as essential mediators in this process. T-cell epitopes are short peptide fragments derived from tumor antigens that are presented on the cell surface, where they trigger a specific immune response. As cancer immunotherapy continues to advance, accurately distinguishing immunogenic epitopes from the vast array of other peptide fragments in an antigen has become increasingly important.

2 Dataset

The primary dataset for this project is obtained from the Immune Epitope Database (IEDB), a publicly available resource that curates experimentally validated epitope data from the peer-reviewed literature. The IEDB team compiles this information by manually extracting and verifying epitope details, ensuring that each entry meets quality standards. For our project, we will focus on human ~~cancer~~ T-cell epitopes. The key variables from IEDB include:

- **Epitope Sequence:** The amino acid string representing the epitope.
- **MHC Restriction:** Whether the epitope is restricted by a specific MHC molecule. Either class I or class II.
- **Source Molecule IRI** Links to the UniProt entry for the antigen that contains the epitope.
- **Predicted Binding Affinity:** The predicted binding affinity of the epitope to the MHC molecule.

In addition to the IEDB data, we will obtain full protein sequences of tumor antigens from the UniProt database. This will allow us to extract non-epitope peptides by identifying segments of the antigen that do not overlap with known epitopes. From both IEDB and UniProt, we

will engineer additional features such as peptide length, amino acid composition, and average hydrophobicity.

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). *The Immune Epitope Database (IEDB): 2018 update* [Data set]. *Nucleic Acids Research, 47*(D1), D339–D343. https://doi.org/10.1093/nar/gky1006. Retrieved from http://www.iedb.org

3 Data Acquisition and Processing

The primary dataset can be downloaded from the IEDB. Appending the full sequences from UniProt is not as straightforward. Some sequences in the IEDB dataset has a feature that contains a URL to the full sequence in UniProt. We will use the `requests` library in python to access the URL and download the full sequence for each epitope. For comparison, I will extract sequences from the UniProt database that do not contain any epitopes.

Key processing steps include:

Data Cleaning:

- Filter the IEDB data to retain only entries corresponding to human cancer T-cell epitopes.
- Remove duplicate records and address any missing or inconsistent values in key variables.

Data Integration:

- Match each epitope with its corresponding full antigen sequence from UniProt.
- Generate negative samples by extracting non-epitope peptides from these full antigen sequences, ensuring the sampled peptides are of matching lengths.

Feature Engineering:

- Compute sequence-based features such as peptide length and amino acid composition.
- Calculate biochemical properties such as average hydrophobicity using the Kyte-Doolittle scale.

4 Research Questions and Methodology

1. What Distinguishes an epitope from any other peptide in an antigen? To answer this question I will use existing features and engineer new features of the peptide sequences. For now, the features I plan to engineer are epitope length, amino acid composition, and average hydrophobicity. Using these features I will perform exploratory data analysis to visualize and understand the characteristics of epitopes and compare with non-epitopes.
2. Can we predict whether a given peptide is an Epitope? To answer this question I will use the features engineered in the previous question to train a model to predict whether a given peptide is an epitope. I will assess the model's performance using accuracy, ROC curve, AUC, and precision and recall.
3. Are there meaningful clusters among epitopes based on their properties? To answer this question I will perform clustering on the epitopes. I will use the elbow method to determine the optimal number of clusters and then use the KMeans algorithm to cluster the epitopes. I will then use the PCA algorithm to reduce the dimensionality of the data and visualize the clusters.

5 Work plan

Week 4 (2/10 - 2/16):

- Data tidying and recoding (4 hours)
- Question 2 (4 hours).]

Week 5 (2/17 - 2/23):

- Aquire data
  - IEDB download (instant)
  - UniProt full sequences (6 hours)
- Data tidying (1 hour)
- Feature engineering (2 hours)
- Generate negative sample data (1 hours)
- Begin RQ1

Week 6 (2/24 - 3/2):

- Week 5 progress report (1 hours)
- Clean up directory and file names (1 hour)
- Complete RQ1 (4 hours)
- Begin modeling (2 hours)
- Implement model evaluations (2 hours)

- Optimize models (5 hours)
- Create Presentation

Week 7 (3/3 - 3/9):

- Week 7 progress report (1 hour)
- Finalize RQ2 (4 hours)
- Presentation prep and practice (4 hours)
- Begin clustering (2 hours)
- Implement clustering evaluations (2 hours)

Week 8 (3/10 - 3/16): *Presentations given on Wed-Thu 3/12-3/13.*

- Week 8 progress report (1 hour)
- Finalize RQ3 (4 hours)
- Write blog post intro (2 hour)
- Poster prep (4 hours)
- Presentation peer review (1.5 hours)

Week 9 (3/24 - 3/30): *Poster Draft 1 due Monday morning 3/24 at 9am. Poster Draft 2 due Sunday night 3/30.*

- Week 9 progress report (1 hour)
- Peer feedback (2 hours)
- Poster revisions (1.5 hours)

Week 10 (3/31 - 4/6): *Final Poster due Sunday 4/6.* - Week 10 progress report (1 hour)

- Peer feedback (1.5 hours)
- Poster revisions (2 hours)

Week 11 (4/7 - 4/13):

- Week 11 progress report (1 hour)
- Write blog post data collection (2 hour)

Week 12 (4/14 - 4/20):

- Week 12 progress report (1 hour)

Week 13 (4/21 - 4/27): *Blog post draft 1 due Sunday night 4/28.*

- Week 13 progress report (1 hour)

[All project work should be done by the end of this

week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/28 - 5/4):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/5 - 5/8): *Final blog post due Tues 5/7. Blog post read-throughs during final exam slot, Thursday May 8th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]

6 References

[The bibliography will automatically get generated. Any sources you

cite in the document will be included. Other entries in the `.bib` file

will not be included.]