**Graph Neural Networks for Immunogenicity Prediction Using Voronoi-Based Protein Representations**  
Marco A. Benavides – mab2511

**Background Introduction and Aims:**

The ability to predict immunogenicity of tumor and bacterial proteins is crucial for advancing cancer immunotherapy and vaccine development. Traditional immunogenicity predictions rely on sequence-based features, but structural properties also play a significant role in determining a protein’s ability to elicit an immune response. Inspired by the paper *BionoiNet: Ligand-Binding Site Classification with Off-the-Shelf Deep Neural Networks*, this project aims to explore Voronoi-based geometric representations of protein structures for immunogenicity classification.

The main objectives of this study are:

1. **Develop a Voronoi-based model** for classifying tumor proteins as immunogenic or non-immunogenic.
2. **Extend to multi-class classification**, incorporating bacterial immunogenic and non-immunogenic proteins.
3. **Evaluate Graph Neural Networks (GNNs)** for processing structural Voronoi representations and compare performance with Convolutional Neural Networks (CNNs) and transformer-based NLP models.
4. **Explore a multi-modal approach**, integrating both structural and sequence-based features to enhance classification accuracy.

**Dataset:**

* The dataset includes 546 immunogenic tumor proteins and 548 non-immunogenic tumor proteins from the Tumor Immunogens Database.
* For multi-class classification, additional bacterial immunogenic and non-immunogenic proteins will be incorporated from *Bacterial Immunogenicity Prediction by Machine Learning Methods*.
* 3D protein structures are obtained from the Protein Data Bank (PDB), and missing structures are predicted using AlphaFold.
* Voronoi diagrams are computed from atomic coordinates to represent protein surfaces and interaction patterns.
* Due to the limited size of the dataset, **data augmentation** techniques will be applied to improve model generalization and mitigate overfitting.

**Workflow:**

1. **Data Preprocessing:**
   * Collect 3D protein structures from PDB.
   * Predict missing structures using AlphaFold.
   * Generate Voronoi diagrams for each protein (example shown below).
2. **Model Development:**
   * Train a Graph Neural Network (GNN) on Voronoi-based representations.
   * Compare with CNNs for Voronoi image classification.
   * If necessary, incorporate amino acid sequences using a transformer-based NLP model.
3. **Evaluation Metrics:**
   * Use cross-validation to mitigate overfitting.
   * Assess accuracy, precision, recall, and F1 score.
   * A diagram of a complex network

     AI-generated content may be incorrect.Perform comparative analysis between GNNs, CNNs, and transformer-based models.

Figure 1. Graphic workflow

**Literature Review and References:**

1. Öztürk, H., Özgür, A., & Ozkirimli, E. (2021). *BionoiNet: Ligand-Binding Site Classification with Off-the-Shelf Deep Neural Networks*. Bioinformatics.
2. Lundegaard, C., et al. (2011). *Bacterial Immunogenicity Prediction by Machine Learning Methods*. PLOS Computational Biology.
3. Doytchinova, I. A., & Flower, D. R. (2007). *VaxiJen: A Server for Prediction of Protective Antigens, Tumor Antigens and Subunit Vaccines*. BMC Bioinformatics, 8:4.
4. Jumper, J., et al. (2021). *Highly Accurate Protein Structure Prediction with AlphaFold*. Nature.
5. Satorras, V. G., & Welling, M. (2021). *E(n)-Equivariant Graph Neural Networks*. NeurIPS. [GitHub Repository](https://github.com/vgsatorras/egnn)