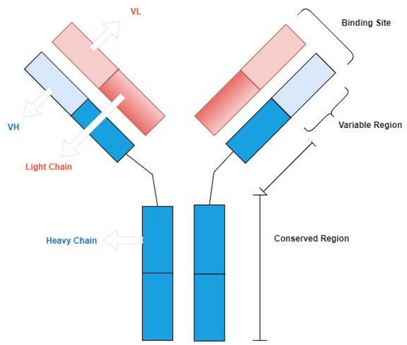
Developing an antibody language model for generating missing amino acid residues to complete partial BCR sequences

Sonal Sujit Prabhu, MS Computer Science
Mentor: Dr. Heewook Lee, Assistant Professor, SCAI
Ira A. Fulton School of Engineering

Biodesign Institute
Arizona State University

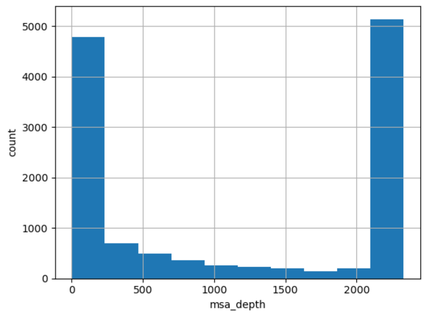
Research Question:
The goal of this project is to build a novel antibody language model for completing partial B cell receptor (BCR) sequences, addressing challenges in immune repertoire reconstruction from RNA-seq data. This model contributes to efficiently profiling therapeutic immune cells, aiding cancer treatment development and infectious disease research

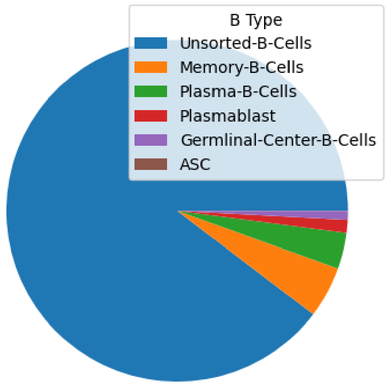
Background:
- Antibodies (BCR) contain heavy and light chains. Located at the ends are the heavy chain variable region (VH) and the light chain variable region (VL), which paired together form the antigen binding site
- Immune receptors from tumor-infiltrating T and B cells hold therapeutic potential.
- Profiling immune receptors via targeted immune repertoire sequencing (TCRs and BCRs) is costly and consumes tissue samples; RNA-seq is also used but may yield partial receptor sequences due to sequencing coverage fluctuations.



Data Filtering and Distribution:
- We extract the unpaired heavy and light chain sequences data from Observed Antibody Space database
- The following filtering steps are undertaken:

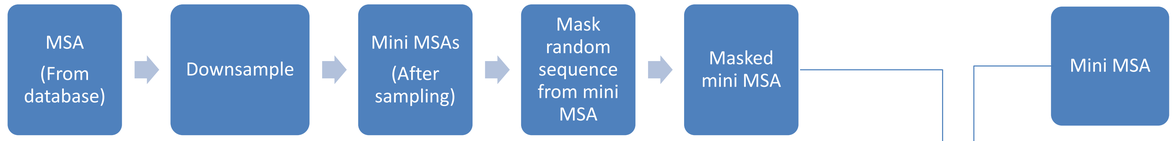
1. We use only human samples for extracting the sequences
2. Exclude Naïve B cell samples for now
3. Downsample large repertoire sequences



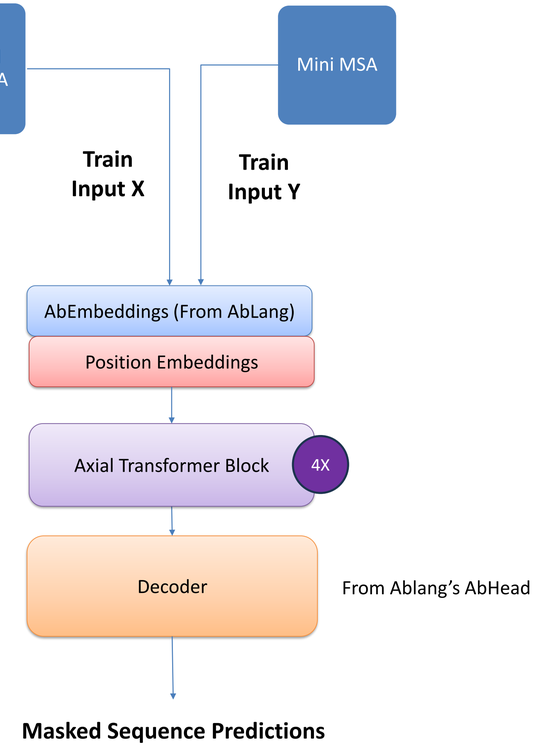


B-cell type distribution

Transformations for Training Data:



Model Framework:
- The idea is taken from AlphaFold's **[1]** EvoFormer embedding architecture, which utilizes an multiple sequence alignment (MSA) of related protein sequences
- We treat BCRs as related protein sequences as B cells undergo clonal expansion, somatic hypermutation and selection to generate a population of related BCRs (antibodies)
- We use a transformer like model that learns contextual patterns within each sequence as well as across related sequences in MSA.
- We use standard attention layer **[2]** across each sequence in MSA and also use an axial attention **[3]** to capture the contextual pattern across each BCR and but also capture relatedness across similar BCRs.
- We add a decoder block for the model to learn how to predict the amino acid token from embeddings learned by the model.



Future Work:
- The training of the above model is still under progress. We aim to bring down the loss of the model and allow for better prediction of the masked sequences

References:

- Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021)
- Ashish Vaswani, Shazeer, and et al. 2017. Attention is all you need. In I. Guyon and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc (2017)
- Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)

Acknowledgements:
I would like to thank Pengfei Zhang (PhD Student, SCAI) for mentoring me throughout the project

MORE
Masters Opportunity for Research in Engineering

Ira A. Fulton Schools of Engineering
Arizona State University