**Minutes**

24/02/23

* MMseqs2 (<https://github.com/soedinglab/mmseqs2/wiki>)
  + This sequence clustering is a less computationally expensive way to generate PSSMs. It is faster and more scalable than HHblits.
    - Pros: Can be used as a replacement for HHblits (in some cases). It offers faster and more scalable sequence searches and clustering.
    - Cons: It offers less sensitive homology detection and alignment than HHblits, therefore HHblits would be important if a sensitive homology detection is very important.
  + Discussed the commands for the workflow of MMseqs2 given an initial FASTA file. I will execute these for this coming week.
* Reviewed data leakage between datasets:
  + Looked at my processing method to separate seeping sequences and prevent any data leakage by moving sequences between datasets.
    - Processing was done using the cluster tsv generated with: ‘mmseqs easy-cluster ALL.fasta clusterRes2 tmp --min-seq-id 0.3 -c 0.8 --cov-mode 1 > mmseqs\_cluster.out’
  + We could also use less of the data by taking the sequence which is flagged as a homologue and ignoring the other ones related to it. This lets us use important sequences which can be the ancestor of these other sequences.
* Training, validation and test datasets:
  + Cross-validation is used often when datasets are very small.
  + The alternative approach is what I have set up with defined train/validation/test splits.
  + Common approach is 80/20 creating a train/test. Then with the training set taking a 80/20 split creating a train/validation.
    - With my current approach I have 80/20 train/test, and 75/25 train/validation, due to splitting by 60/20/20.
  + After hyperparameters have been trained using the training and validation datasets it is important to train the model using both these datasets, then evaluating the model’s performance on the test dataset.
  + Finally, the test dataset is often included in the learning before final deployment.
* Dissertation
  + Looked over section headings to make sure the correct sections were being included.
  + Focusing on relevant material and content is important. There are many alternative approaches to protein disorder prediction, but it would be best to discuss things that I have went on to work with and used.

Goals for this week:

* Generate PSSMs using MMseqs2.
* Finish background section of dissertation.
* Do all top-level writing of design and implementation sections.
* Automate movement of sequences between training, validation and test datasets as required.