**Minutes**

27/01/23

* Looked over RNN for predicting protein disorder:
  + Model appears to be fine
  + RNN handles the hidden state internally. I was confused about passing something extra to the model, but the batched sequence is all.
* Looked over training loop with RNN model:
  + Is also fine.
  + I’m using einsum, but could always transpose dimensions 1 and 2.
  + To note with my squeezing of batches:
    - Usually loss function takes in a batch of predictions (e.g., [class A, class B, class B,…] of perhaps 16 items. Then, makes the optimiser step.
    - As I’m batching with 1 and have a list of labels [[0,0,0,…,0,1,1,…,1,0....]], then I torch squeeze to produce a batch of labels. Here, each amino acid is treated as a prediction of class either ordered or disordered.
  + I should check output of the loss computation.
    - To make sure the correct ordered/disordered labels are being chosen.
      * I can experiment with weights on each log probability, with customisation of BCELoss. The weighting is often calculated using the ratio of labels. This helps emphasise predictions of disordered (unbalanced) label.
      * For example, if there are 1/5 1’s (disordered amino acids) and 4/5 0’s (ordered amino acids), then I could multiply the probability given for a 1 (disorder) appearing by 4 and a 0 (order) appearing by 1.
* Discussed encodings:
  + There are protein language models that create tokens/encodings of each amino acid in the sequence. These can help learning.
  + They would replace the current one-hot encoding implementation.
  + An example from an initial search after the meeting is Facebook’s ESM protein language model.
* Discussed PSI-BLAST.
  + Produces a PSSM matrix – this input should also improve learning.
  + To use this, it takes FASTA data (accession number and sequence) as input. It can also take in TSV data. I will need to read on how it works.
* Discussed dissertation:
  + Should include a short background of what proteins are before discussion about IDPs.
  + Research projects follow a more methodology section instead of the analysis/design/implementation so investigate structure of dissertation.

Goals for this week:

* To finish most of the introduction and background sections.
* Improve models by assessing BCELoss weights and input other than one-hot encodings.
* To stick to the plan: to download the CASP datasets and use them with my models.