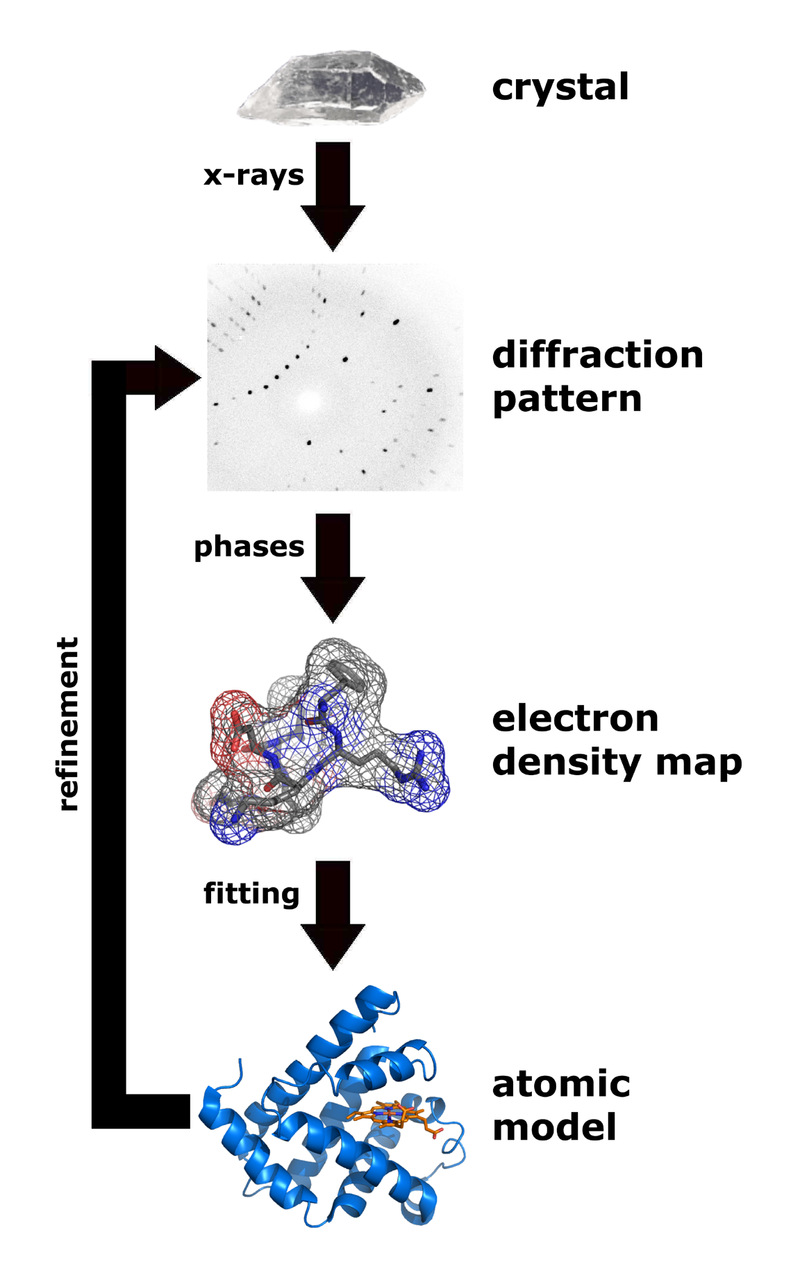
**Minutes**

20/10/22

* Definition/terminology clarifications from today:
  + IDRS are regions/sections of disorder within a disordered protein. So a IDP may have 1, 2, or more IDRs.
  + Definition clarification: residue is an amino acid, regions are sequences of amino acids that form a structure.
* Discussed single method approaches vs consensus method approaches (approaches that use several single method approaches and then takes a consensus of how to classify the protein). The DISOPRED3 does this. This will take a lot longer than single method approaches as we are effectively running 3 single method approaches. This also means that it will likely be better because it is using the single method approach plus extra help.
* Discussed PDB and GraphQL. Seen that data is not straightforward to find using these. Planning to still look at this a little bit this week while finding datasets I will continue with.
* Looked at DisProt database: <https://disprot.org/download> . So, data from here shows annotated disordered proteins and IDRs.
* DisProt has TSV and FASTA data. TSV data can be manipulated using Python (pandas), and can extract relevant data from FASTA using PyTorch DataLoader.
* DisProt file has ~5500-6000 entries which will be sufficient and good for training, validating and testing the model.
* DisProt file has the protein ID. This can be used on PDB and UniProt to also search the protein. DisProt file contains the start and end positions of the disordered protein and the disordered sequence. We can see on the other site, PDB, that searching this protein gives the full protein sequence with this disordered region greyed out and not visualised.
* Looked at X-ray crystallography – where the small protein crystal is used to create the diffraction pattern – forming the electron density map (where higher more dense values are stronger and less wobbly) - this is then turned into the atomic model. <https://en.wikipedia.org/wiki/X-ray_crystallography>
  + 
* Can still set up agent where a goal is to be able to retrain on updated DisProt dataset, and other datasets considered.
* Discussed DISOPRED papers – items I’d annotated and was confused about.
  + PSI-Blast – Finds similarities amongst the database using sequence matching.
  + Discovered a better competitor compared to PSI-Blast – called HHblits. Would be used as an input to a DNN.
  + PSSM created from PSI-Blast. PSSM is linearly scaled to [0.0, 1.0] - normalisation. This means that large numbers do not have a more powerful weighting and does not trip up the machine learning algorithm. E.g, when buying a house bedrooms range 1-5, but a garden can be 200-300 square feet, yet bedrooms usually more improtant for the price than big garden.
  + Asked about positive and negative training sets. Positive sets are the disordered proteins. The negative training sets is an ordered protein – this is not what we are trying to identify. Could be wise to train the model on this because then the model won’t be very confused when it sees an ordered protein.
* CASP – can use this for benchmarking my solution. Then I can directly compare my solution with other recognised methods that have already been tested on CASP (unseen dataset).
* Personal goal is to complete the deep learning by Christmas.

Goals for this week:

* PyTorch practise.
* Look at DataLoader in PyTorch
* Browse research papers
* Aim to build the dataset. Dataset is the priority this week.
* Look at all info DisProt file gives me. Consider what can be classified and used for learning.