Benedikt Hopf
Alireza Ketabdari

1	2	Σ

Exercise Sheet Nr. 6

(Deadline February 03, 2022)

Problem 1

(a) A 2D contact map is a matrix/table where every row and every column correspond to one amino acid of the sequence. The value at location ij then represents weather the two amino acids i and j have contact (i.e. are touching each other). Basically if one thinks of a protein as a undirected graph, where the nodes are the amino acids and the edges, indicate, that two amino acids are touching each other, then the 2D contact map is simply the adjacency matrix of the graph.

(b) Amino acids can be encoded by a 25 dimensional vector. This vector is set up as follows:

- 20 dimensions are used to one-hot encode the type of amino acid (of which there are 20, that are actually used).
- Three more dimensions are used to encode (one-hot) the secondary structure, that the amino acid is in (i.e. α-helix, β-sheet or coil).
- Two dimensions are used to specify, weather the amino acid is on the outside (exposed) or on the inside (buried) of the final protein (also one-hot).

Obviously since there are only 20 amino acids one could technically simply encode an amino acid as a 20-dimensional one-hot vector and be done, but obviously the other information are also helpful, so they are used as well.

(c) Since the accuracy of such models is often not that high, often only the *L/5 top scored predicted contact pairs* are used. L here is the length of the amino acid, so $L/5$ means, that only the best $L/5$ amino acid contacts are used for evaluation. Therefore a model that perfectly predicts on these 20% and incorrectly on everything else would achiece a perfect score.

(d) One way to measure the dependence of random variables is to test, how well one can be predicted, given the other. This is obviously not trivial, since there are lots of ways to try to predict one variable from the other, so maybe they are independent, but maybe also the model used for prediction is not good enough.

An easy way to measure dependence (however not perfectly) is to assume an affine model. This can be easily done using the correltion coefficient given as

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]\mathbb{E}[(Y-\mathbb{E}[Y])^2]}} \in [-1,1]$$

where $|Corr(X,Y)| = 1$ indicates perfect affine dependence, and independence leads to $Corr(X,Y) = 0$. However $Corr(X,Y) = 0$ does not necessarily imply independence. This is because the correlation coefficient only considers linear dependence, so non-linear independence might not be visible.

In any machine learning task, the output must be dependent on the input, otherwise learning is obviously impossible. So technically the performance of for example a neural network, could be used as a measure of dependence (it will perform poorly, if there is little to no dependence), but obviously, the model might also perform poorly, because it is a bad model despite there being some dependence.

(e) Pooling in neural networks is used to reduce the resolution/size of a hidden layer, in order to make the problem computationally more tractable. There are several types of pooling, like average-pooling, but max-pooling is most common. Max-pooling is done, by selecting the largest value from some specified region of the input and discarding the rest. How large that region is and how many regions are looked at is a hyperparameter that can be set. So an example could look like that:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \overset{max-pooling}{\longrightarrow} 4$$

Here there is only one region that is looked at which has a size of 2×2. If the input was larger e.g. 8×8 and we would still look at 2×2 regions, then the result would be 2×2, and thus size would have been decreased.

Obviously max-pooling is a procedure with information loss, since only the maximum element is retained and everything else is discarded, this way one could for example get

$$mp\left(\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\right) = 4 = mp\left(\begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}\right)$$

where mp denotes the max-pooling operation.

However, this loss of information is not too bad since an activation in a convolutional network (where pooling is usually used) can be seen as the presence of some feature (say an edge or a circle) do max-pooling keeps all the most important features and only looses the ones that are weaker/not present, which is not much of a loss.

(f) For viruses there are two infection routes:

- *Direct contact:* This is where there is direct contact between the already infected host (who transmits the virus) and the new host. This can happen is several ways, for example by blood contact or by aerosoles (as it is with COVID-19) where the infected host breathes and exhales particles that include some virus, which can then directly infect the new host.
- *Indirect contact:* This is where there is no direct contact between the two hosts. This can for example happen, if the infected host touches some surface and leaves some residue behind. The new host can then come to that surface later, when the other host is already gone and touch the surface (and thus the infected residue) and get infected that way.

(g) There are two parts to the immune system:

- *Innate Immune System*: This is the evolutionarily older part of the immune system and thus the main part of the immune system for less complex organismns like plants or insects. The main distinguishing factor is, that the innate immune system is innate and thus does not change during the lifetime of an organism. Therefore there is no learning or adapting to diseases.

One important part of the innate immune system are white blood cells. Their job is to take away foreign substances from e.g. the blood, and thus keep the blood clean.

Another innate part of the immune system is the skin which has the obvious job of being a barrier between the body and the outside world.

So the innate immune system is not very specific against the disease, and works against basically anything to some degree. However it can not specialize on substances/viruses/bacteria to better fight them if they have mechanisms to defend against the innate immune system[1].

This is where the adaptive immune system comes in:

- *Adaptive Immune System*: This is the evolutionarily newer part of the immune system. The main difference here is that (as the name says) it is able to adapt to previously unknown threats. This is why for example vaccines only work on the adaptive immune system since that is able to learn from it.

  If there is some patogene (e.g. a virus) present, this first activates T-cells. These then transform into T helper cells and T killer cells. The helpers (which are activated by HLA class II molecules) can activate the killer cells as well as B cells. The actual immune response is then done by these B cells and T killer cells.

  The B cells produce antibodies which bind to the patogene and by that render it useless (e.g. by binding to the spike protein of a virus and therefore inhibiting its function). T killer cells look for HLA class I molecules and then kill the (human) cells which have produced them. This makes sense since these cells are then infected by the virus and by killing them the reproduction of the pathogenes can be stopped (since for example viruses cannot reproduce by themselves).

Problem 2

Predicting the structure of that protein results in a contact map (Figure 1) as well as a 3D structure (Figure 2). These look very different from the experimental results (Figure 3). The experimental structure looks a bit like an actual spike as it is a somewhat compact structure with a wider portion on the bottom and slimmer at the top (if oriented as if Figure 3). The prediction is a much more elongated structure with specifically one α-helix extending out quit far (lower right in Figure 2). If one was to fold that α-helix upwards (again in that image) then it would have some similarity to the experimental results, even though not too much still.

There are probably two reasons for that. One is, that these predictions just are not that precise (except for AlphaFold2) and the second one is, that a part of the protein sequence is missing (since one can only provide the first 1000 amino acids, but the acual protein has closer to 1300.

[1] Information mostly taken from `https://en.wikipedia.org/wiki/Innate_immune_system`

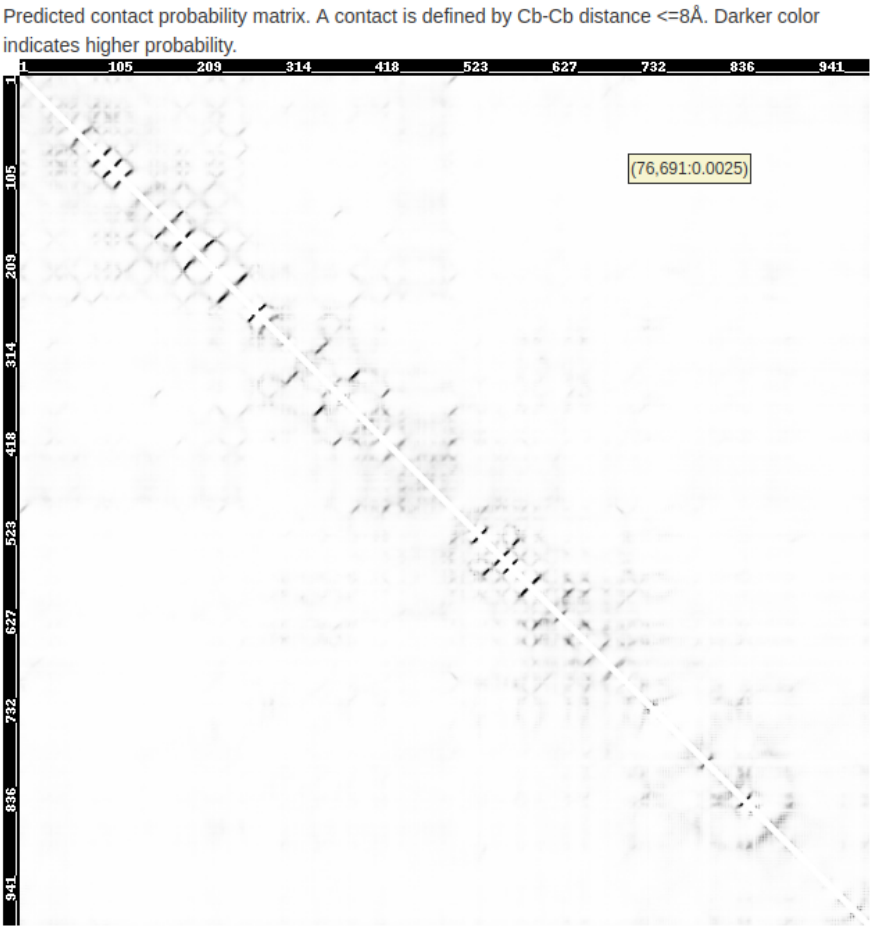

Figure 1: Contact map as created by RaptorX

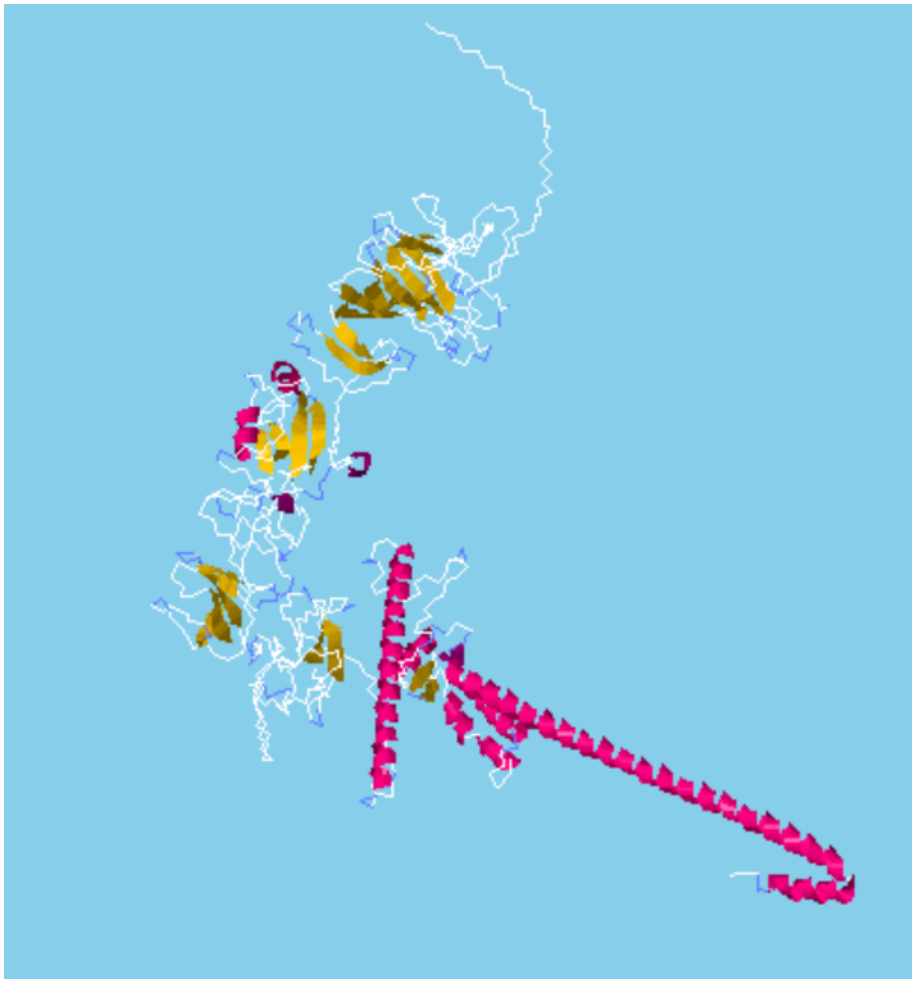

Figure 2: 3D structure as created by RaptorX

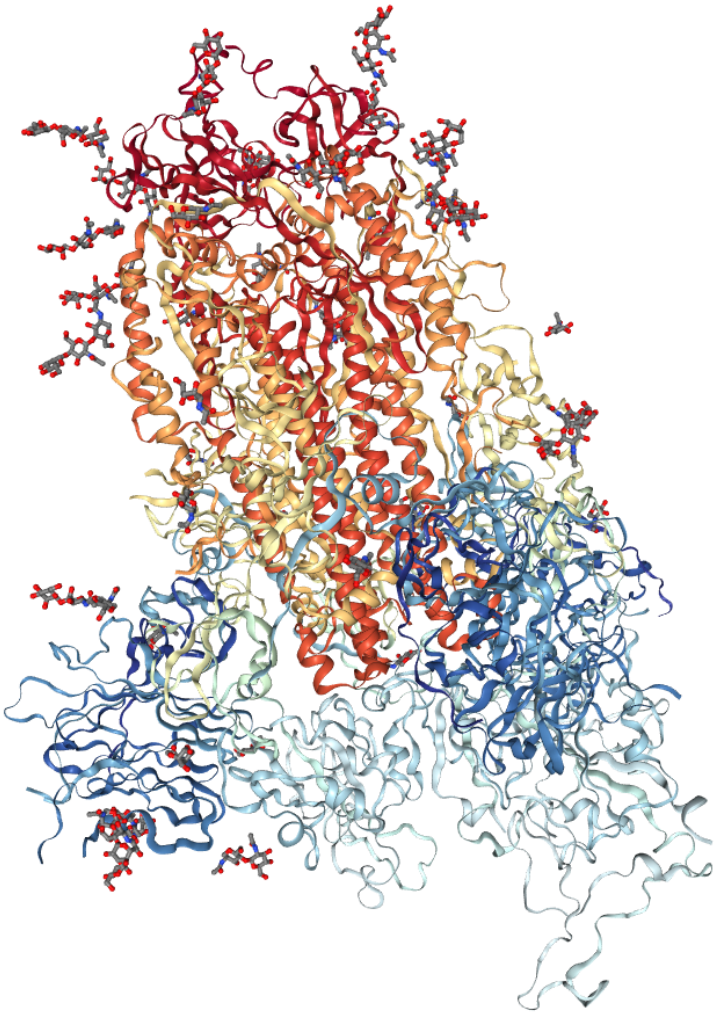

Figure 3: Experimental structure as provided by Protein data bank