3D Data Processing in Structural Biology - Assignment 4

Noam Sabag Bickel

Shay Bergman

Embedding layer/size

1. Changing the chosen layer, changes the representation the model is using, which means that each baseline is based only on the layer and thus we get different results for each layer chosen. The main pattern we saw was that for low/early layers we got high AUC, then a large drop-off once we moved to using mid-level layers and then a smaller jump up when using high/late layers.

2. No, In our case the best model(based on test AUC) was one with an embedding layer of 640 (Although we couldn’t run any larger than 1280 because of RAM insufficiency). The larger models succeeded better on the train set (peak of 0.007 test accuracy), and we speculate that they performed worse on the test set due to overfitting.

Distance baseline

3. The score formula distinguishes classes because it measures whether a peptide is closer to positive or negative examples in the embedding space: a higher score means the peptide is more similar to positives and less similar to negatives, making it likely to be a true positive.

Neural-network classifier

4. The best Test AUC score we got was 0.93, for that we used
Embedding size=640, Embedding layer=17, test size=0.2, batch size=64, epochs=50, lr=1e-3, hidden dim=128 and dropout=0.4.
5. a. We could add to the ESM embedding vector complementary features like predicted secondary structure, or structure-based scores such as mean pLDDT or peptide to CRM1 COM distance, to give the network information that embeddings alone might miss.
b. One option is to replace the simple neural network with sequence aware(models like CNNs or transformers, to give more impact to patterns in the data.

Unsupervised structure

6. The classes are not completely separable in 2d but the majority of P samples seems to be on the left(and N on the right), so k means manages to find somewhat meaningful clusters.

AlphaFold COM analysis

7. We got ROC AUC values of 0.47 for COM distance and 0.76 for mean pLDDT, thus we would say that mean pLDDT is a better discriminator as it separates positive samples from negative ones better than the COM distance.