Ex 4 – Answers

Sivan Mor Yosef, Lior Zats, Yair Mahfud, Yuval Cohen

1. Changing the embedding layer significantly affected baseline performance, with lower layers actually outperforming higher ones (layer 9: AUC = 0.91, layer 20: AUC = 0.66, layer 33: AUC = 0.84). This is actually a well-known phenomenon where intermediate layers in protein language models capture better features for downstream tasks than the final layers, which get too specialized for the original training objective.

2. In general, larger models don't always lead to improved performance as they can suffer from overfitting, especially with limited data, and sometimes the added complexity just introduces noise rather than useful signal. However, in our case, increasing embedding size up to 2560 dimensions did lead to nearly perfect performance (AUC = 0.98), though this improvement likely wouldn't continue indefinitely with even larger models.

3. This score compares how far a point is from negative examples versus positive examples. When a point is close to positive examples (small dist_to_pos) and far from negative examples (large dist_to_neg), the score becomes high, indicating a positive prediction. The log1p(x) = log(1+x) transformation helps by handling zero distances safely and compressing large distance differences, making the classifier more robust to outliers.

4.
```
AUC: 0.9824999999999999
```

   The best test-set AUC achieved was 0.98 using embedding size 2560 from layer 9, with a neural network trained for 41 epochs using batch size 128, learning rate 5e-4, hidden dimension 128, and dropout 0.2. This near-perfect performance demonstrates that large ESM embeddings combined with intermediate transformer layers can effectively capture protein-protein interaction patterns.

5. (a) Additional structural inputs could include 3D coordinates or secondary structure elements from PDB files, which would help because protein interactions are fundamentally driven by 3D shape complementarity and binding sites that

sequence alone might miss. However, given our near-perfect AUC of 0.98, these additions might only provide marginal improvements.

(b) Alternative approaches could include graph neural networks (GNNs), which could represent proteins as graphs where amino acids are nodes connected by spatial contacts or sequence proximity, allowing the model to capture 3D structural relationships and long-range dependencies that linear sequence models might miss.

6. The t-SNE plot shows a decent class separation with orange points concentrated more toward the right/upper regions and blue points toward the left/lower areas, though in the middle region there's more overlap between classes while still maintaining some structure with orange points favoring the right-middle and blue points favoring the left-middle areas. K-means fails to find meaningful clusters, simply assigning the upper region to orange and middle/lower regions to blue, which doesn't capture the actual left-right separation patterns visible in the t-SNE visualization.

7. Mean pLDDT is a significantly better discriminator than COM distance, with ROC AUC values of 0.76 versus 0.47 respectively. The COM distance AUC of 0.47 is actually worse than random chance (0.5), indicating that simple geometric distance between protein centers provides no meaningful signal for predicting interactions, while pLDDT captures structural quality information that actually relates to whether proteins are likely to interact.