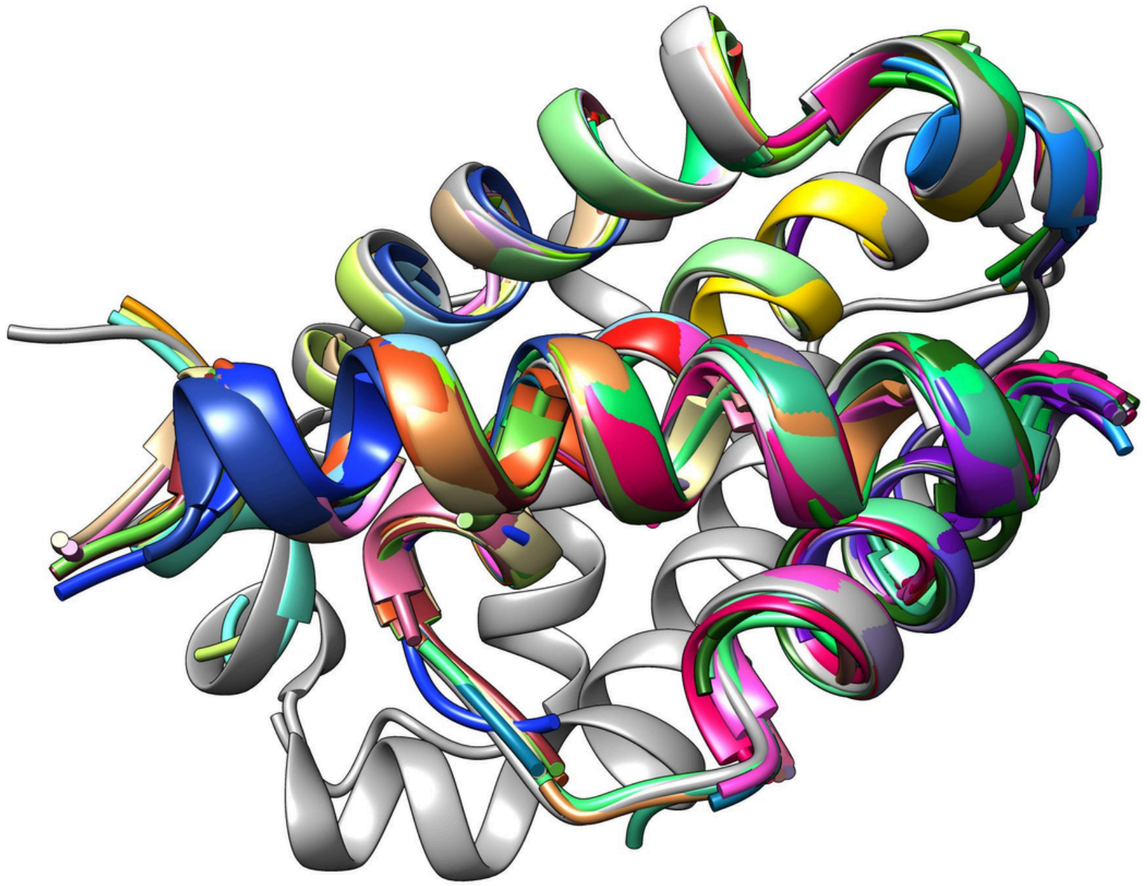
Unsupervised and supervised analysis of protein sequences

Machine Learning Project

Joan Descoubes
Luís Leitão

Protein Structure



- Protein: a sequence of amino acids
- Amino acid : static part (same for all amino acids) + variable part
- Variable part : contains 3 bases made of 4 different elements.
- Overall, 20 different natural amino acids are distinguished.

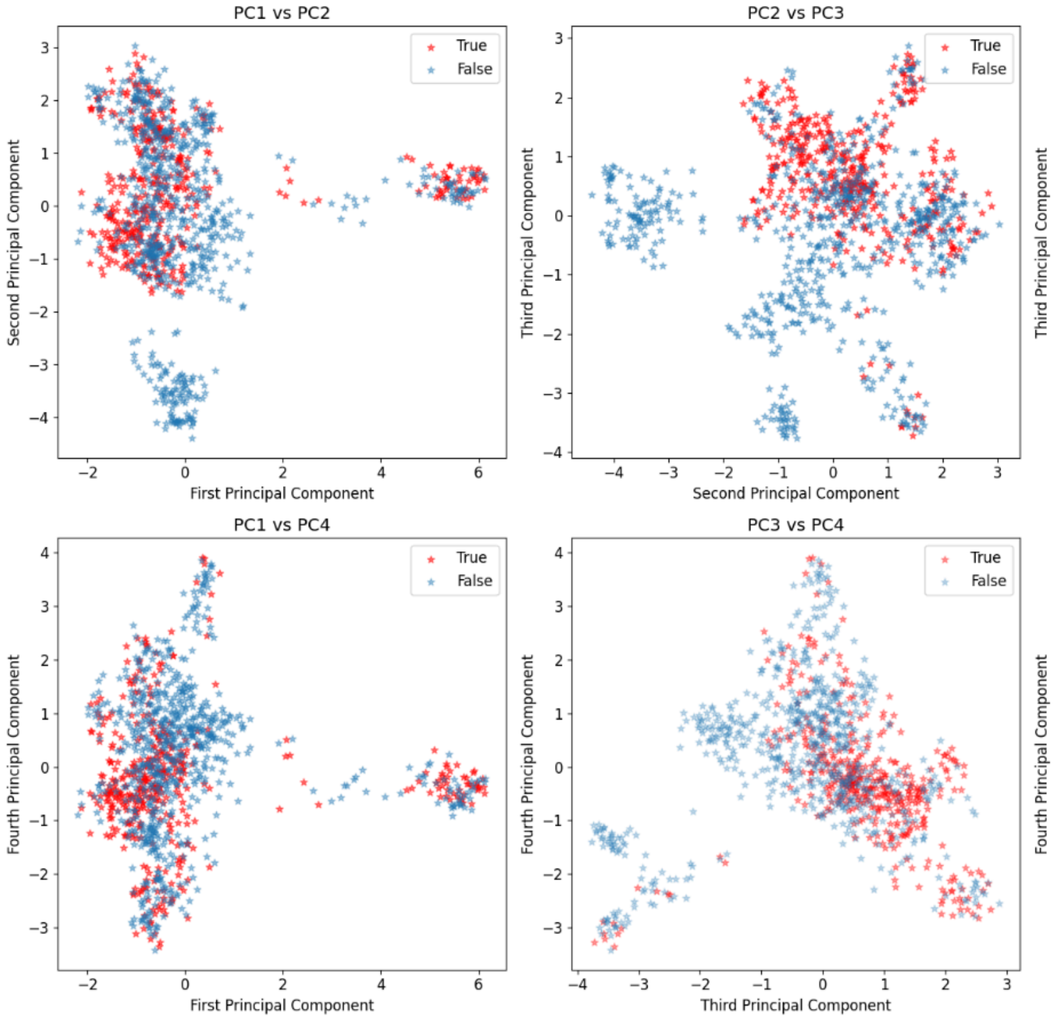
TASK 2 : PCA

TASK 2 : PCA

- Principal component analysis (PCA) : Finds the (orthonormal) directions of maximum variance in the data.

- No rescaling : variance along directions of original data is expected to be informative (proportional to how often amino acids are present at a given position of the protein)

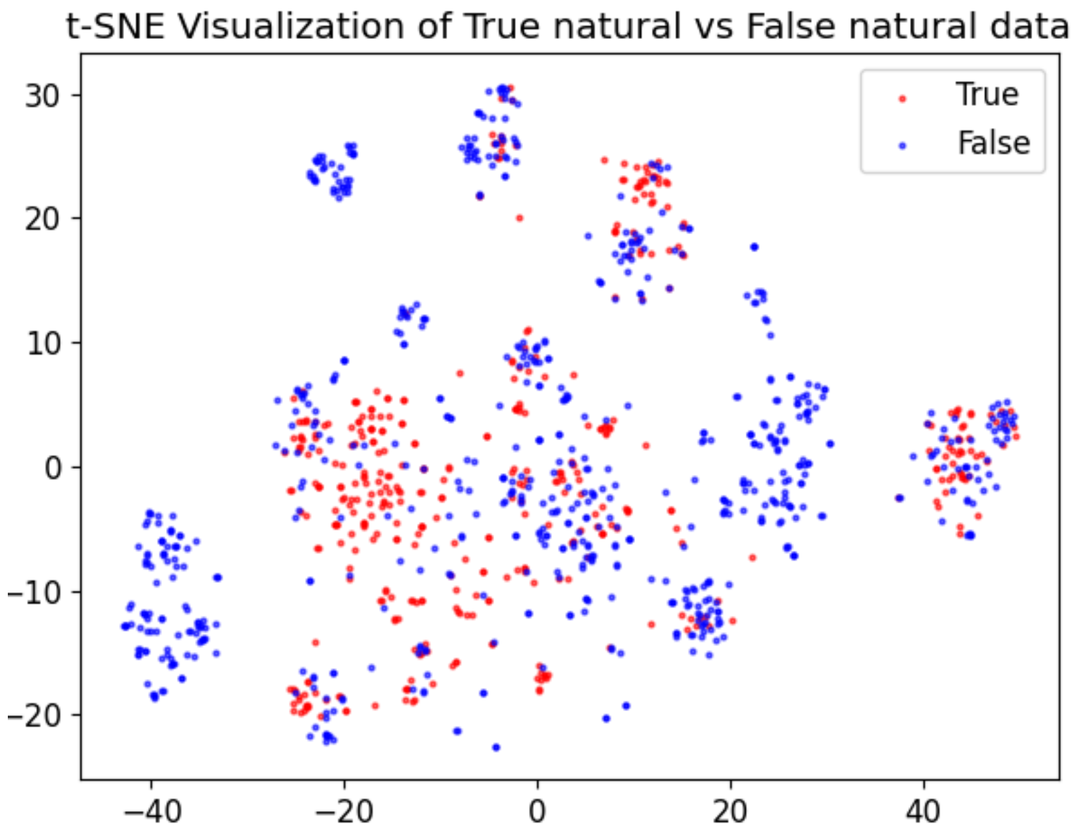
Task 2 : PCA True vs False



- True : the protein is functioning in an experimental screen.

- Presence of clusters dense regions separated by sparse regions.
- True data points occupy a smaller region of space.
- From 4 first principal components True and False data points are not well separated (there is overlap).

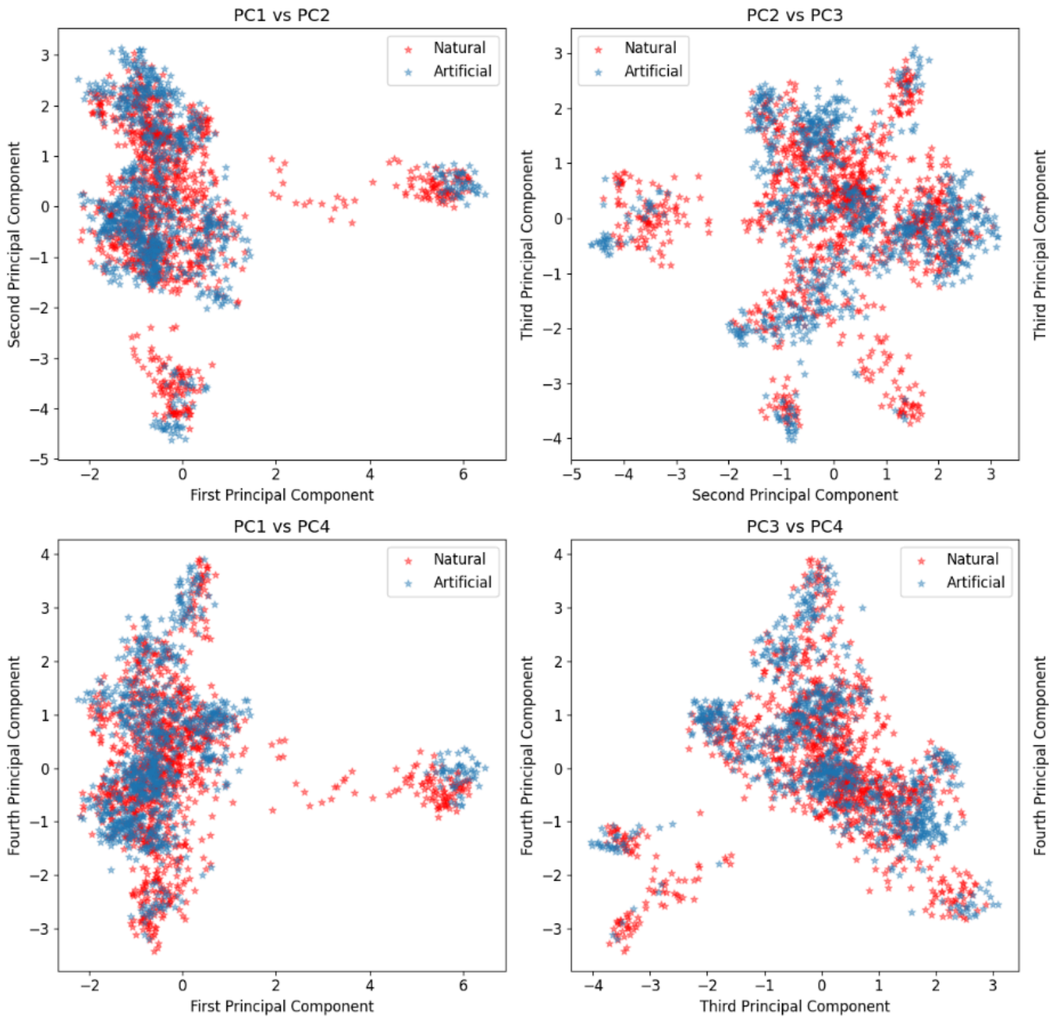
Task 2 : t-SNE



- t-SNE : reduce the dimension through a non linear mapping. Should preserve the order of distances between points.

- Appearence of clusters
- True and False sequences cannot be well separated.
- True data points occupy a smaller region of the overall space.

Task 2 : PCA natural vs artificial

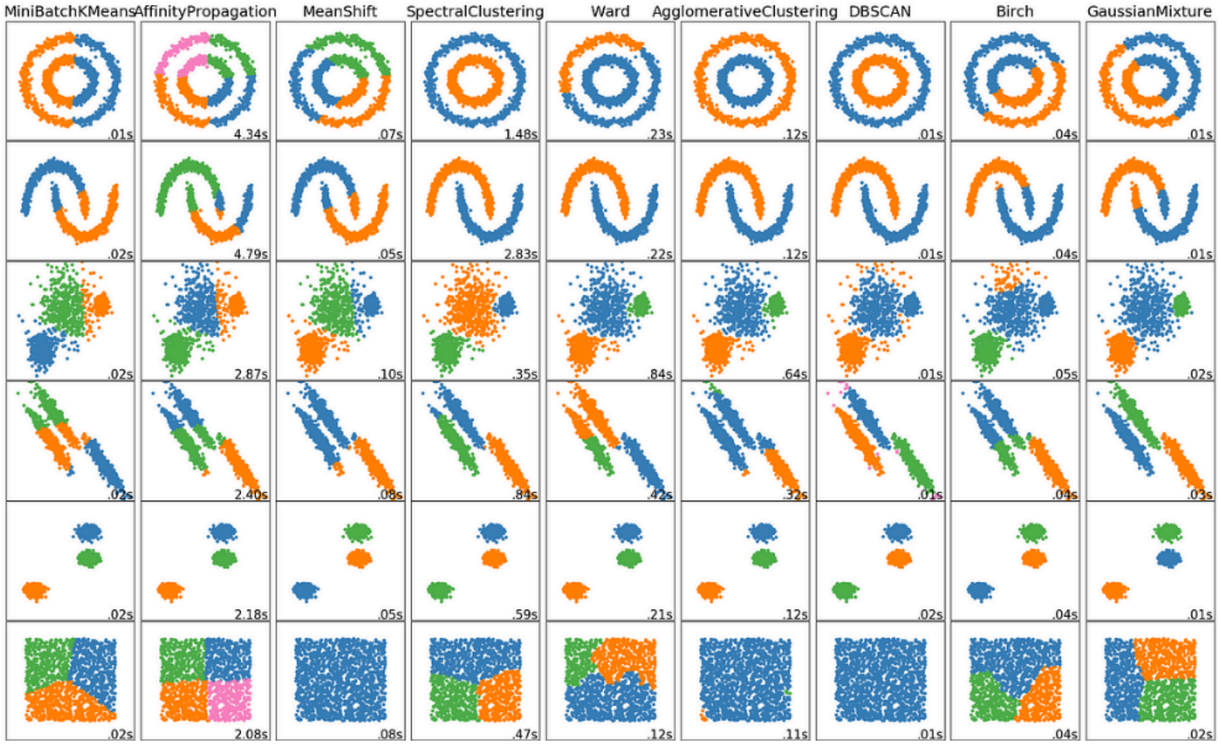


- Artificial sequences : generated by a generative model learned on the natural dataset.

- From 4 first PCs natural and artificial sequences occupy similar regions of space.
- Natural data occupies a slightly broader region of space.

TASK 3 : CLUSTERING

TASK 3 : Clustering algorithm



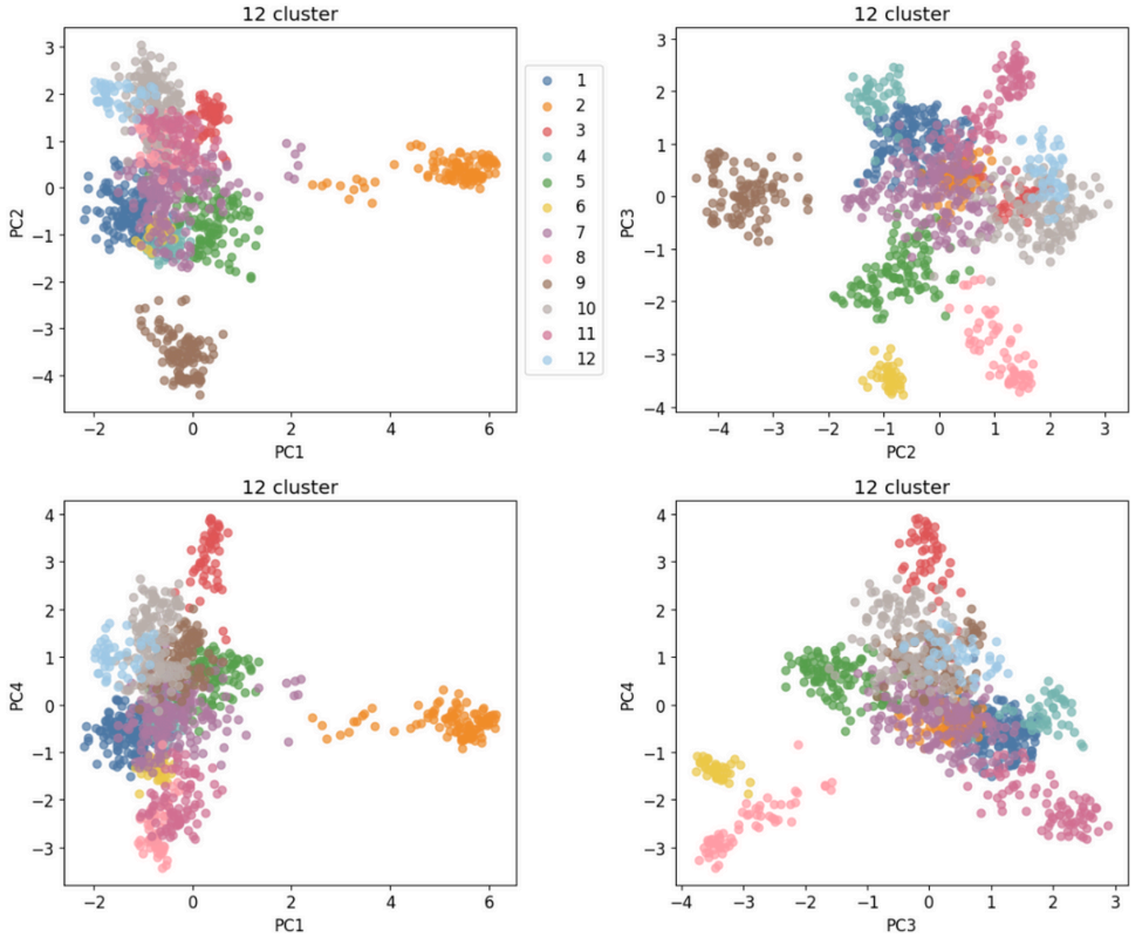
- Structure of the data is best captured by line 3 to 5.
- Gaussian Mixture Model (GMM) seems like the best algorithm for this type of data
- GMM : Clusters are generated using gaussians. The parameters (mean, covariance matrix) are learned by iterating.

TASK 3 : GMM

- Do clustering for 1 to 25 component.
- Find the number of components that minimises the Akaike Information Criterion ($AIC = -2 \cdot \log(P(Data)) + 2 \cdot K$) .
- Find value of the log likelihood after wich the increase is not significant
- Use the average of the two as the number of component for the clustering.

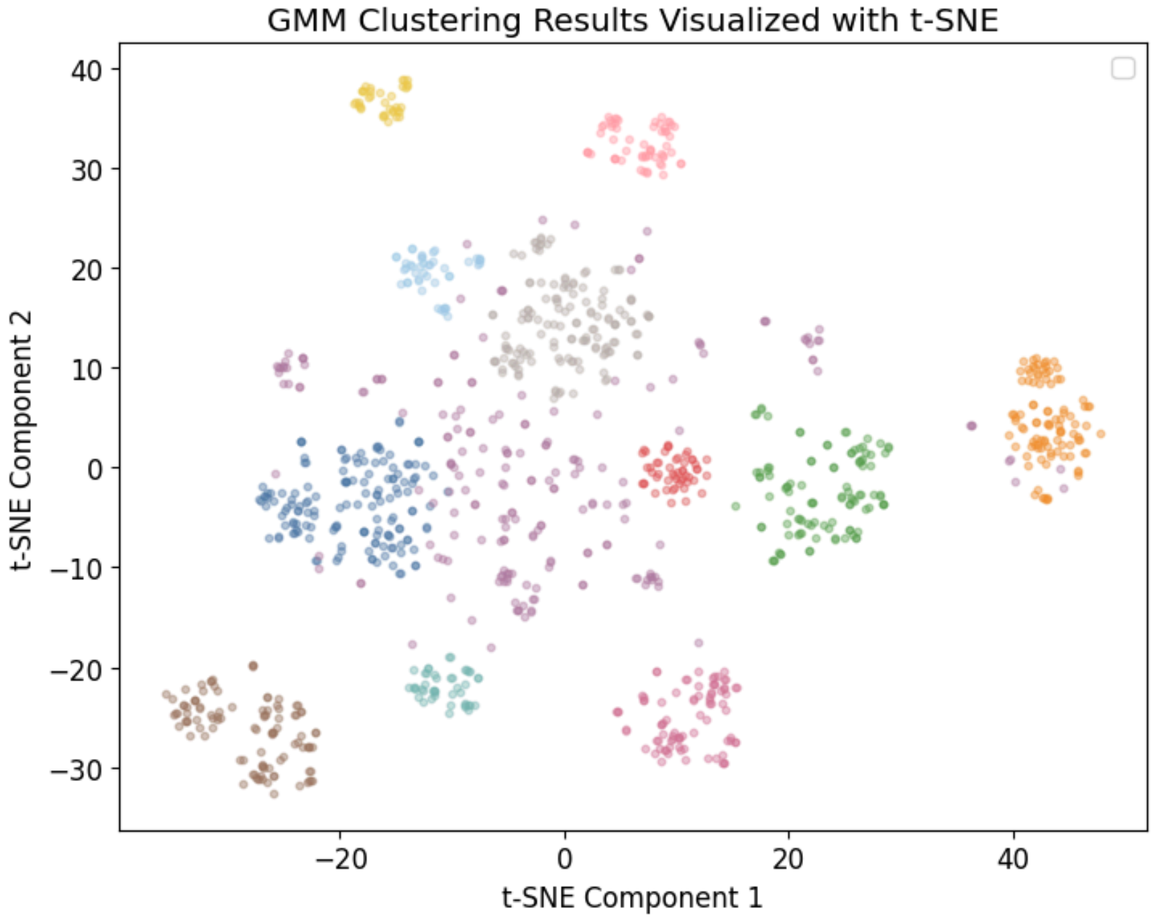
- For the natural dataset 12 components are used and 16 when natural and artificial datasets are merged.
- Clustering is done on 600 dimensional data (with PCA)

TASK 3 : GMM for natural sequences



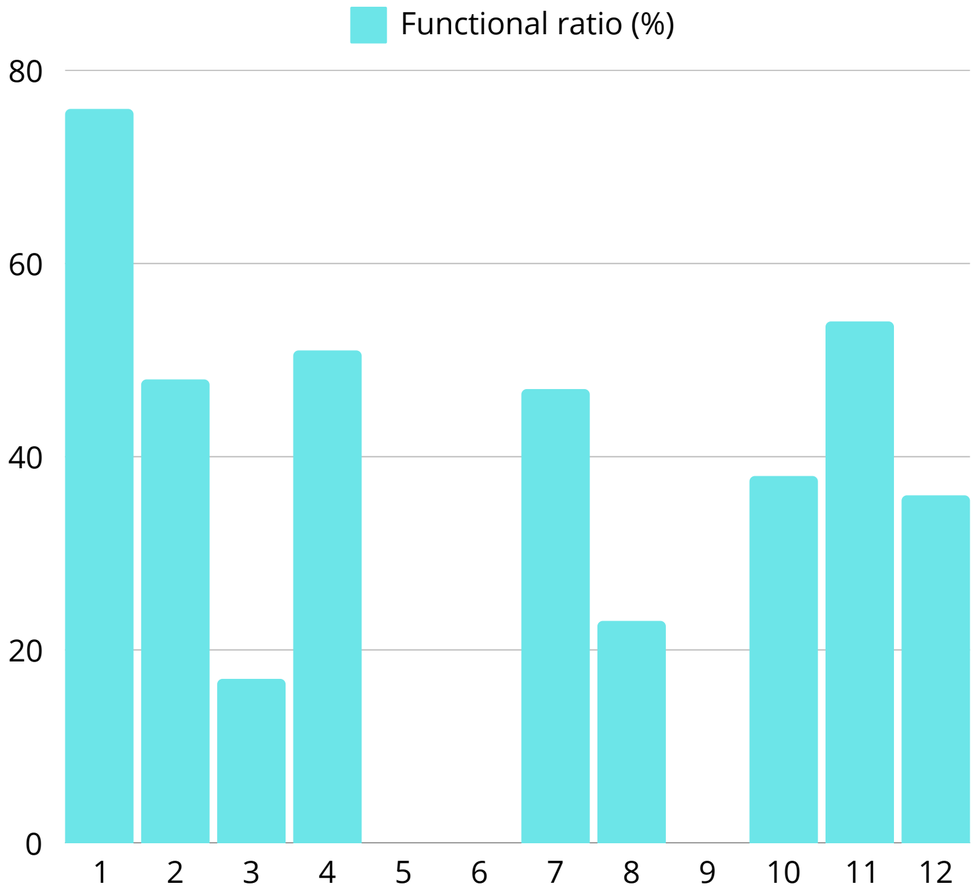
- From PCA of task two most clusters are overlapping
- Except for cluster 2, 6, 8 and 9 clusters cannot be distinguished properly from the first four PCs

TASK 3 : GMM for natural sequences



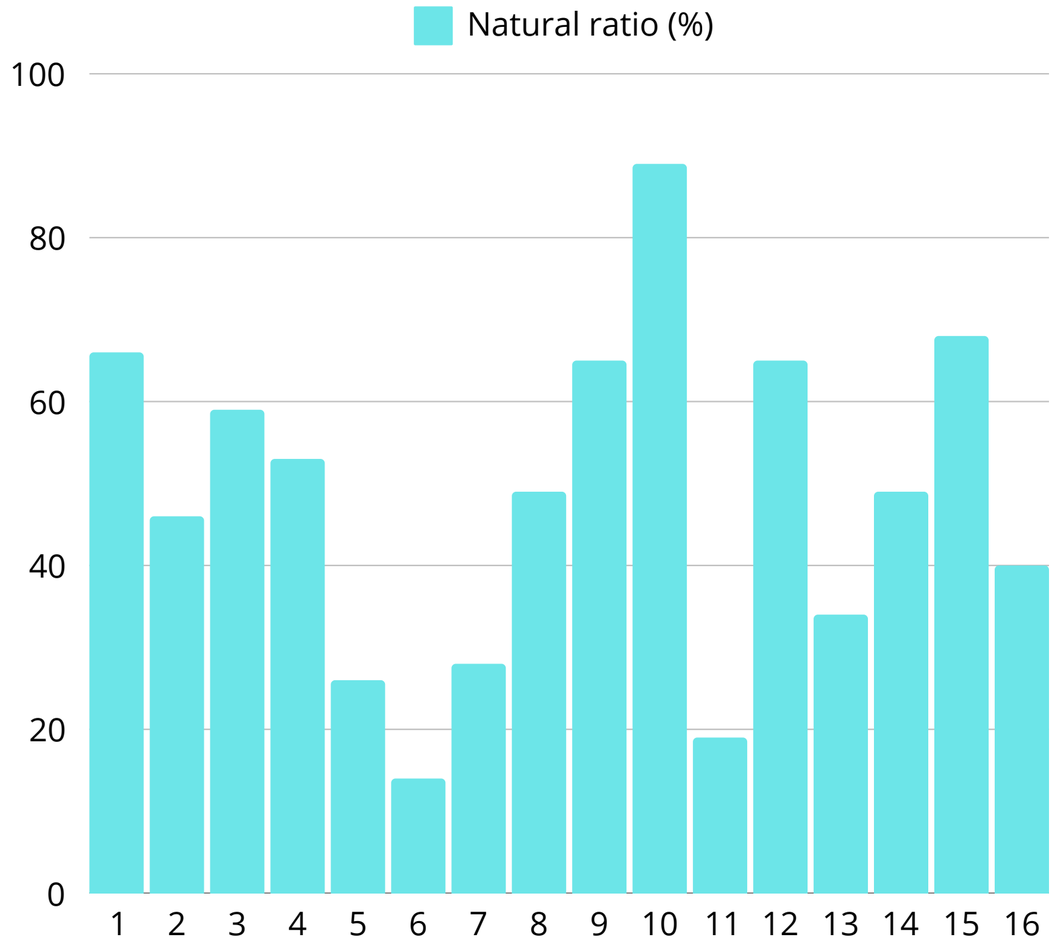
- Using t-SNE allows to visualise the clusters better
- t-SNE and GMM seem coherent with each other
- Clusters are well separated except for number 7 (darkest purple) that covers a large area of t-SNE space with a low density of points.

TASK 3 : GMM for natural sequences



- Functional ratio : $\frac{\#True}{\#True + \#False}$
- Functional and nonfunctional sequences are not separated in different clusters
- 3 clusters are well separated (clusters 5, 6, 9)
- Non functional sequences are present in more clusters

TASK 3 : GMM for all sequences



- The two datasets are not well separated by the clustering.
- The natural ratio is between 14% and 89%.

TASK 4 : Logistic Regression

Task 4: Theory

$$s_i = \boldsymbol{x}_i^T \boldsymbol{w} + b_0 \equiv \mathbf{x}_i^T \mathbf{w}$$

$$P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + \mathrm{e}^{-\mathbf{x}_i^T \boldsymbol{\theta}}},$$

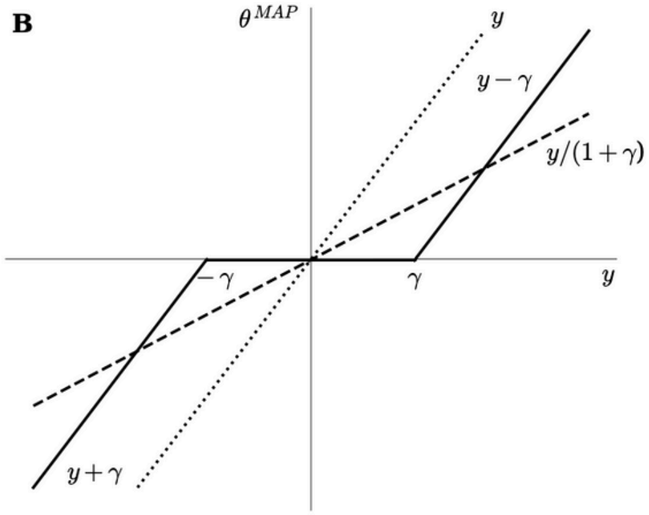
$$P(y_i = 0|\boldsymbol{x}_i, \boldsymbol{\theta}) = 1 - P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\theta})$$

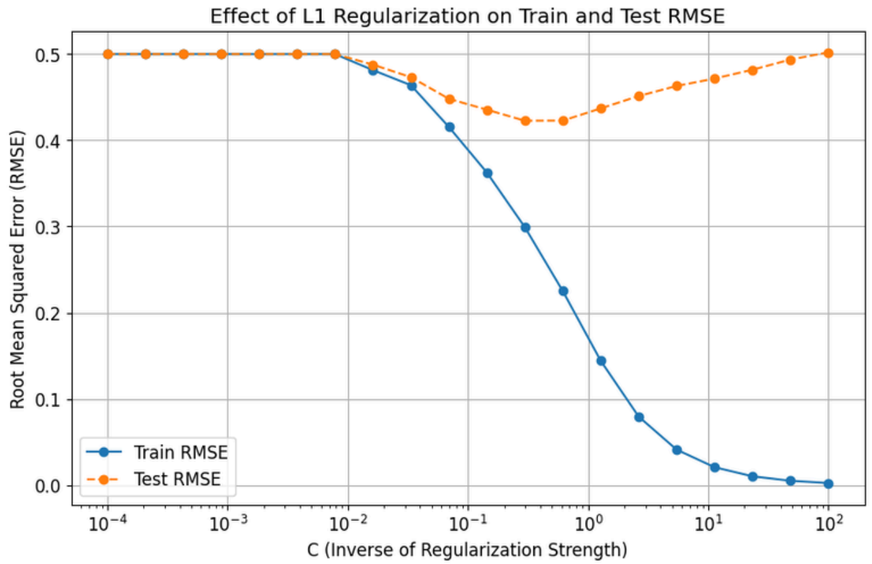
$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

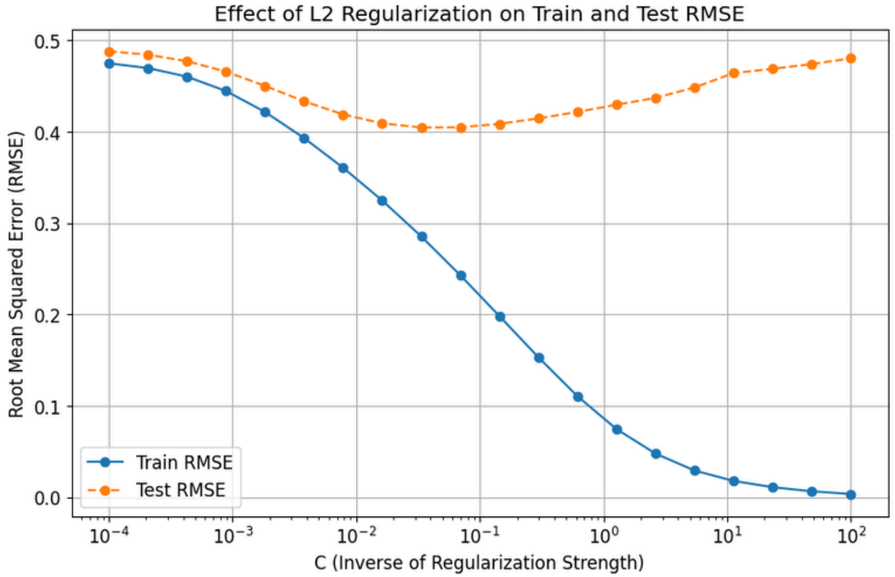
- Goal: predict whether a protein is functional (y=1) or non-functional (y=0);
- Linear combination of the input features weighted by coefficients;
- The sigmoid function transforms s into a probability;
- The model learns the coefficients by minimizing the cross-entropy.

Task 4: Methodology

- Cross Validation;
- L2 regularization better than L1;
- To avoid overfitting, we used C = 0.034.

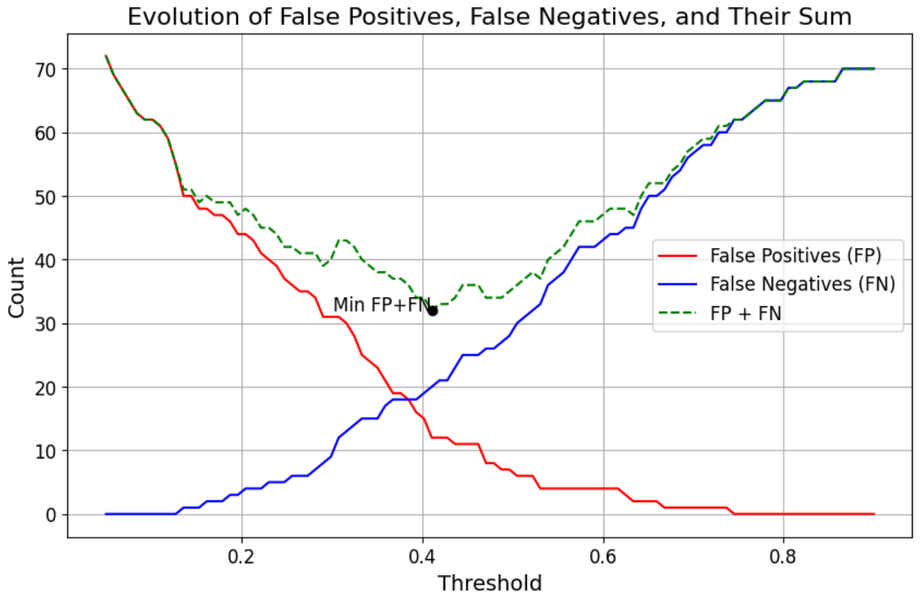






Task 4: Results

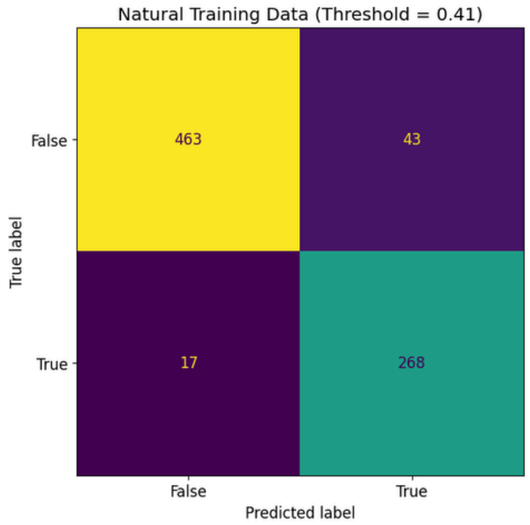
- To minimize the number of total false predictions, we could change the threshold from our hard classifier from 0.5 to 0.41;

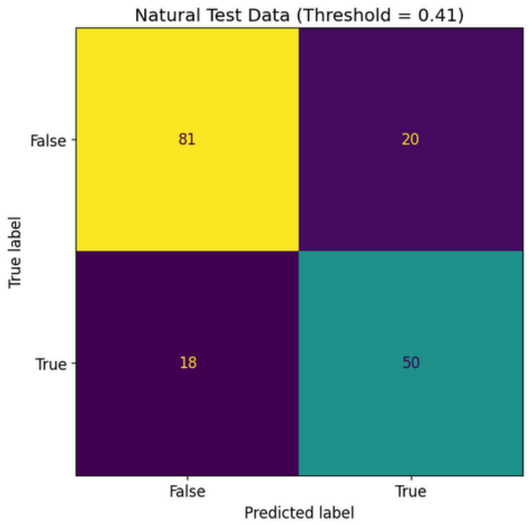


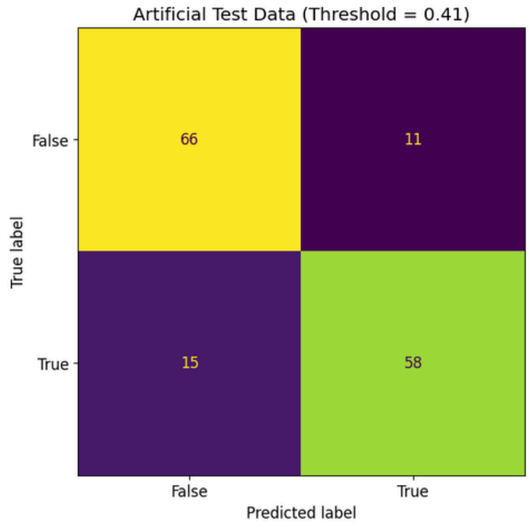
Natural Train data:
FP = 8%
FN = 6%

Natural Test data:
FP = 20%
FN = 26%

Artificial Test data:
FP = 14%
FN = 20%



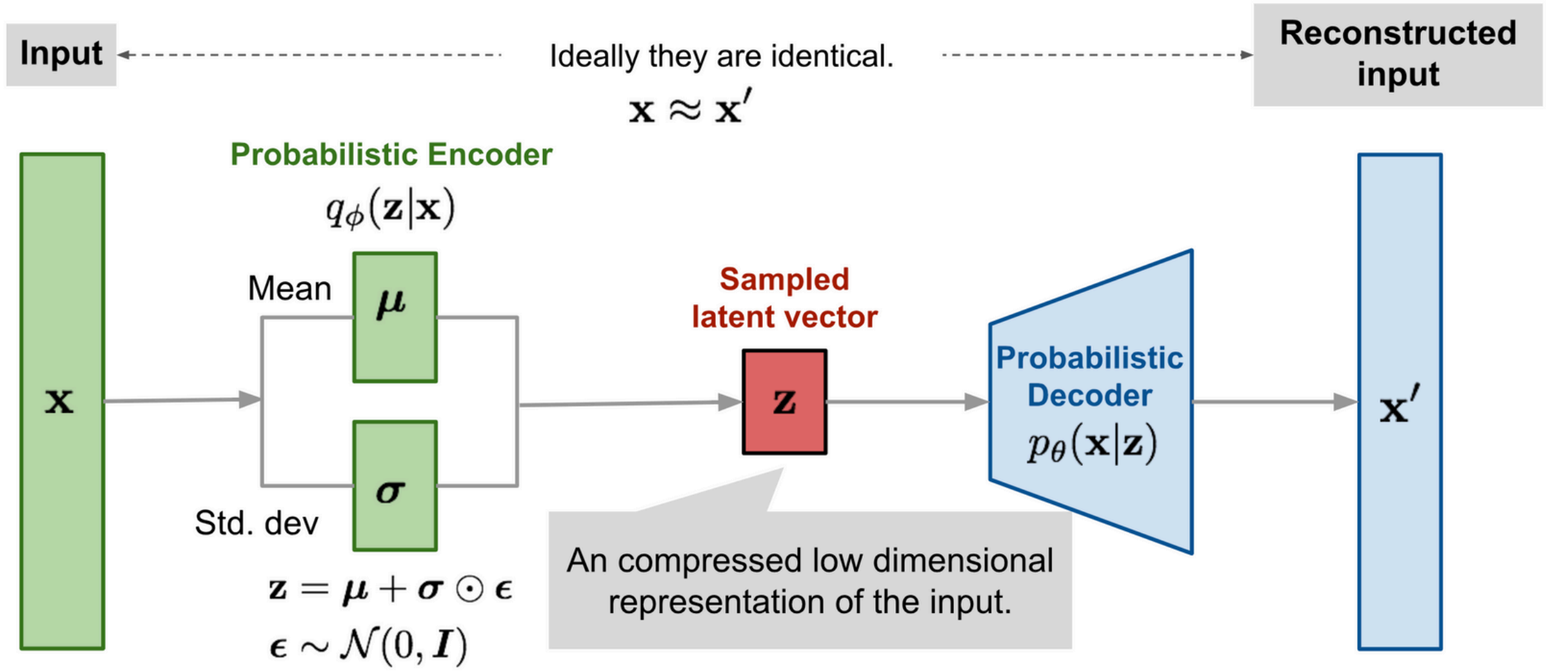




TASK 5 : VAE

Task 5: Theory

- Protein sequences are encoded into a latent vector representation, z, using the posterior distribution, modeled as a Gaussian;
- To ensure differentiability during training, a reparametrization trick is used;
- The decoder reconstructs proteins from the latent vector.
- Random points are sampled from the Gaussian, and the likelihood distribution is used to model the probability of generating a specific protein sequence from z.
- The loss function measures how well the model reconstructs the proteins and is updated each epoch.



Task 5: Methodology

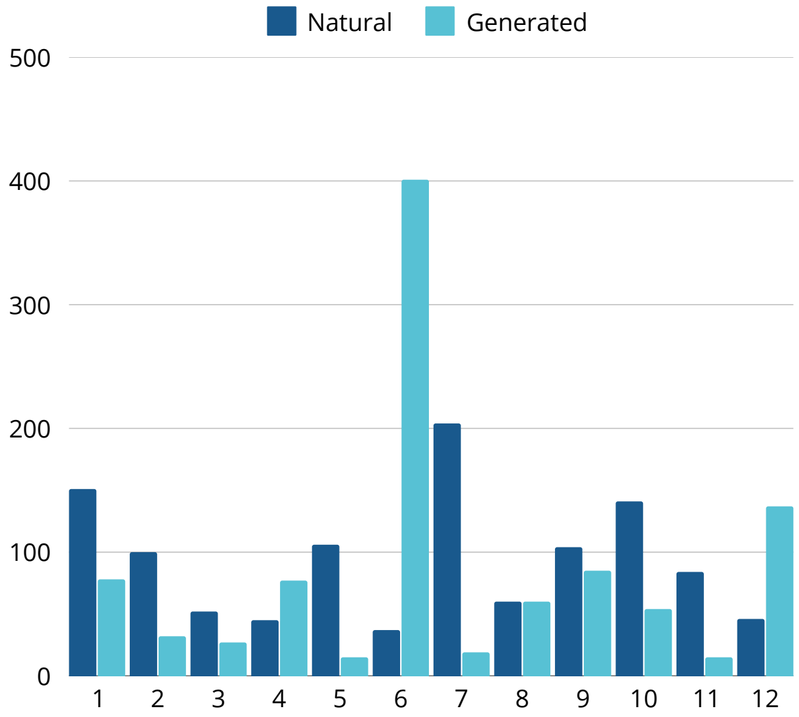
- The model assigns a probability to each element in the amino acid vector using a Sigmoid.
- Using torch.argmax, the model selects the most probable amino acid at each position.
- New hot encoding: amino acid vector with dimension 21!
- Best hyperparameters find through iteration.

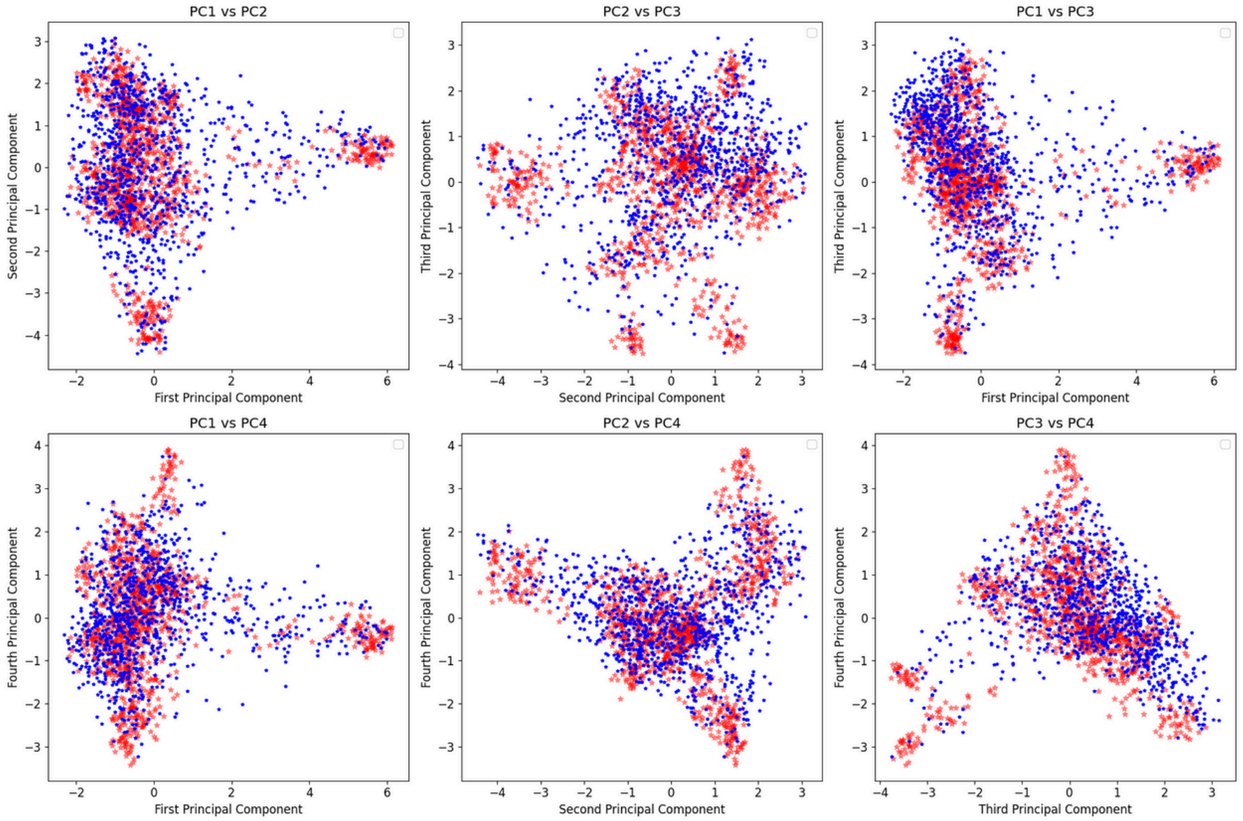
Hyperparameters

- **Latent Vector Dimension: 10** - Determines the size of the encoded representation.
- **Batch Size: 30** - Number of sequences processed in one iteration.
- **Epochs: 60** - Number of training iterations over the dataset.

Task 5: Results

- The projection of the generated data onto the principal components closely mirrors the distribution of the natural training data.
- The sequence generator appears to sample from all clusters.
- The proportions of points sampled from different clusters differ between the natural data and the generated data.





Conclusion

PCA

Clusters can be observed in PCA, True and False sequences are not well separated, natural and artificial datasets seems to occupy the same regions of space

Clustering

Clustering does not allow to separate neither True and False sequences nor artificial and natural ones.

Supervised Learning

Logistic Regression performs better than a random classifier, however for achieving higher accuracy, implementing an ensemble method may be advisable

Unsupervised learning

VAE effectively captures the essential characteristics of the protein sequence space

Thank you for your attention!