DAT405 Assignment 3 – Group 11

Jihad Almahal - (7 hrs)
Oscar Karbin - (7 hrs)

February 7, 2023

1 Show the distribution of phi and psi combinations using:

1.a A scatter plot

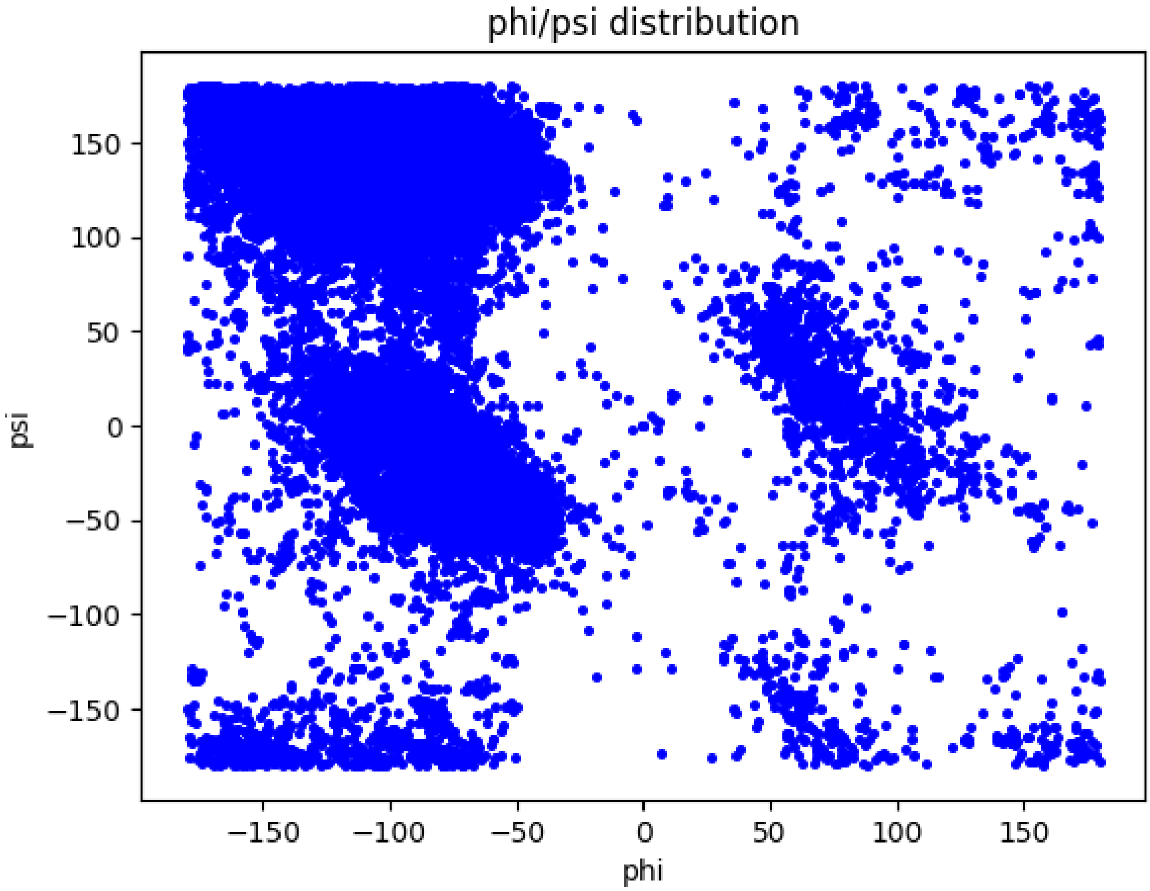

Figure 1: distribution of phi and psi

1.b A 2D histogram

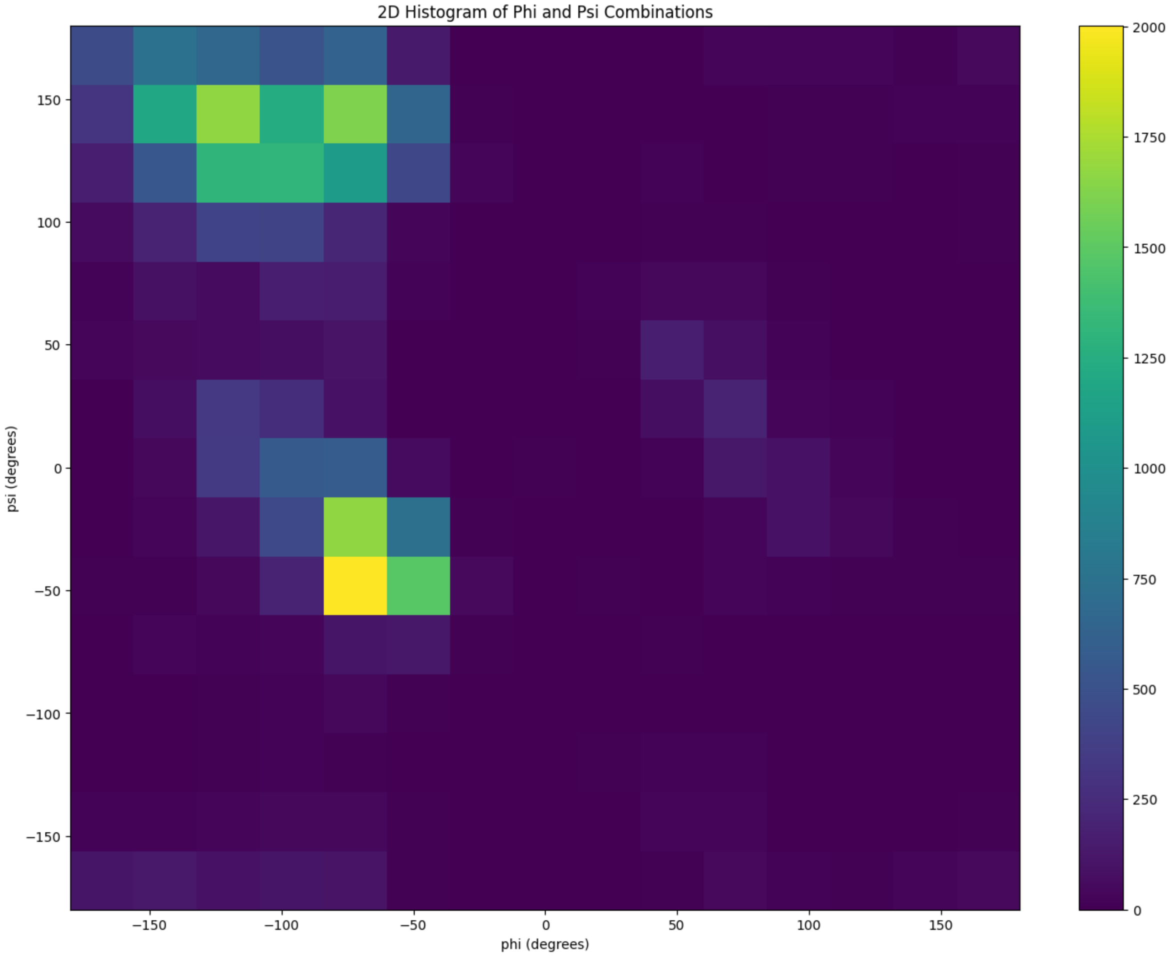

Figure 2: A 2D histogram of phi and psi distribution at 15 bins

There is no fixed rule for the optimal number of bins, but a commonly used rule of thumb is the ”Sturges rule”, which suggests using k=log2(n)+1 bins, where (n) is the number of data points. This generated 15 bins.

2 Use the K-means clustering method to cluster the phi and psi angle combinations in the data file

2.a Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.

K-means is a clustering method for grouping similar data points into clusters. In this task, we used K-means to cluster the phi and psi angle combinations.

To perform K-means clustering in Python, we used the KMeans class from the scikit-learn library. We performed the K-means clustering with K=3.

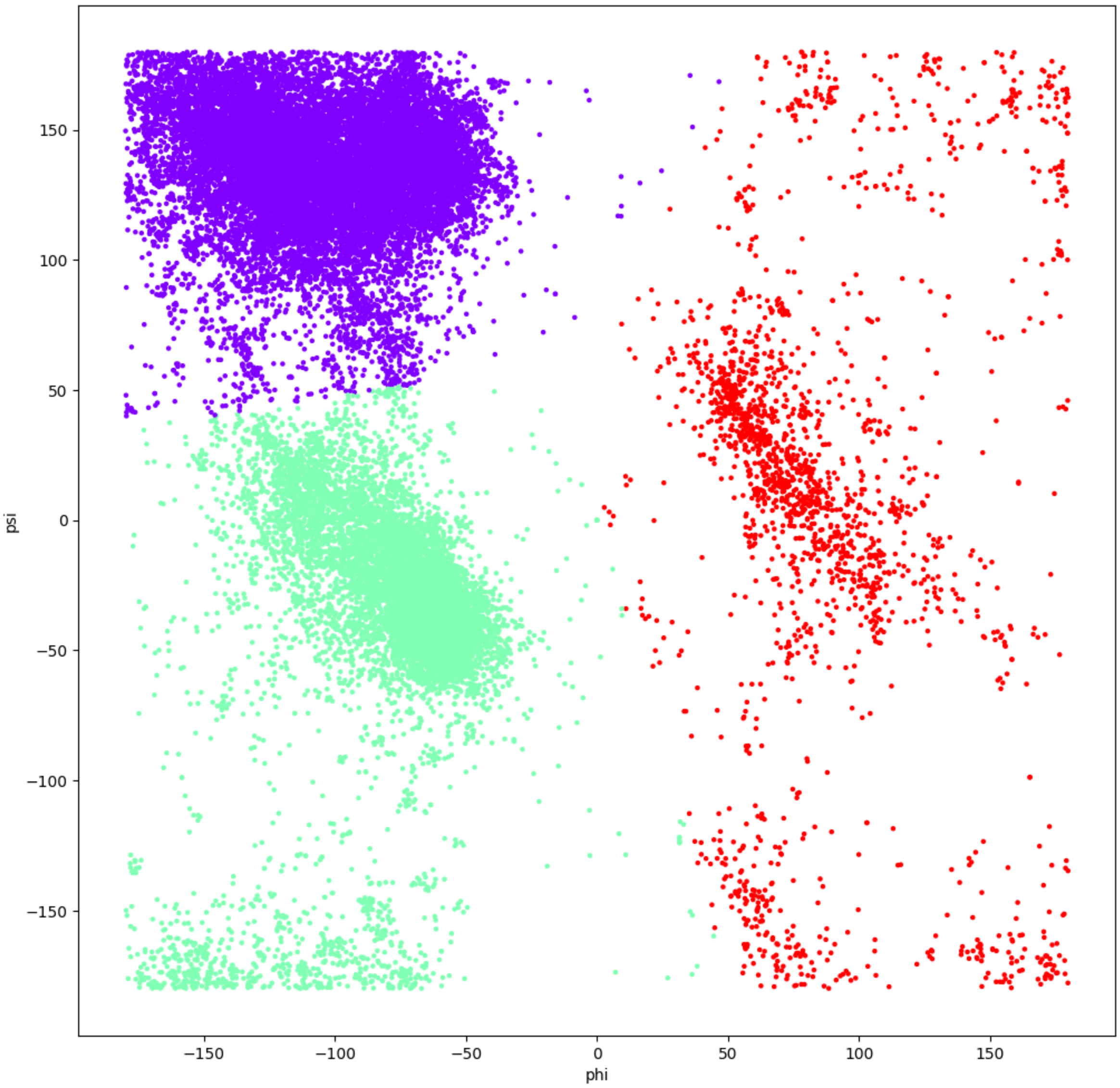

Figure 3: K-means clustering with k=3

To determine an appropriate value of K, we used the elbow method. The elbow method involves plotting the sum of squared distances between data points and their closest cluster center for different values of K and selecting the value of K at which the sum of squared distances starts to decrease at a slower rate. This value is considered to be the optimal number of clusters. Here is the result we got when running the elbow method.

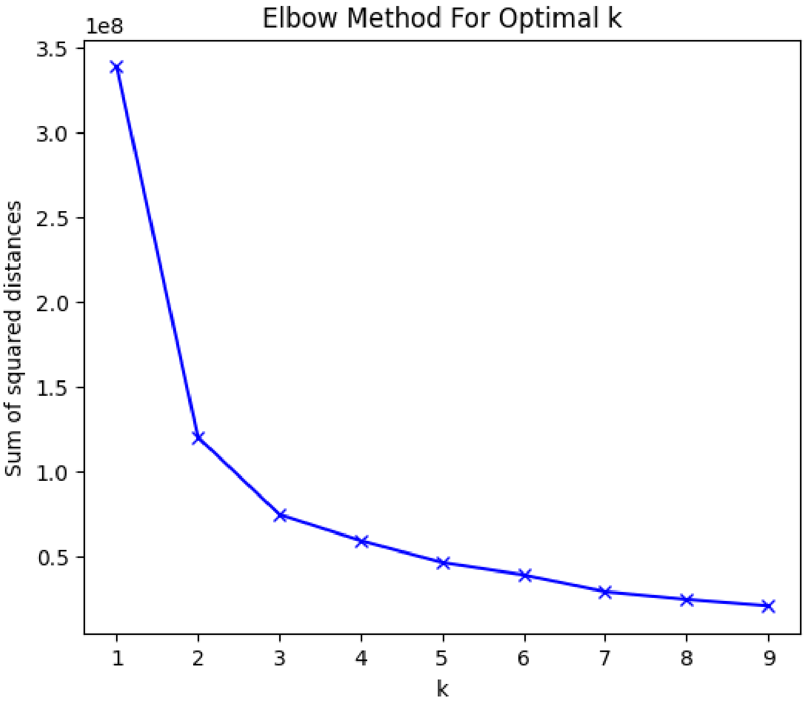

Figure 4: Elbow method for finding suitable K value

The optimal number of clusters is the value of K where the sum of squared distances starts to decrease at a slower rate, which in this case was a K value of 3.

2.b Do the clusters found in part (a) seem reasonable?

Determining whether the clusters found in part (a) are reasonable depends on the interpretation of the data and the desired outcome. It also depends on the value of K that was chosen and the method used to evaluate the clustering results.

One way to evaluate the clustering results is to use silhouette coefficient, which measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates that the clustering results are reasonable. The score we got was a score of 0.67 which considered to be a relatively good score. A score of 0.67 suggests that the objects are relatively well-clustered, but there may still be room for improvement. It is important to keep in mind that the silhouette score is only one method for evaluating the clustering results.

3 Use the DBSCAN method to cluster the phi and psi angle combinations in the data file.

By modulating two parameters, we can use DBSCAN to estimate the number of clusters and identify the noise: epsilon, which is the shortest distance (euclidean distance) and the minimum samples.

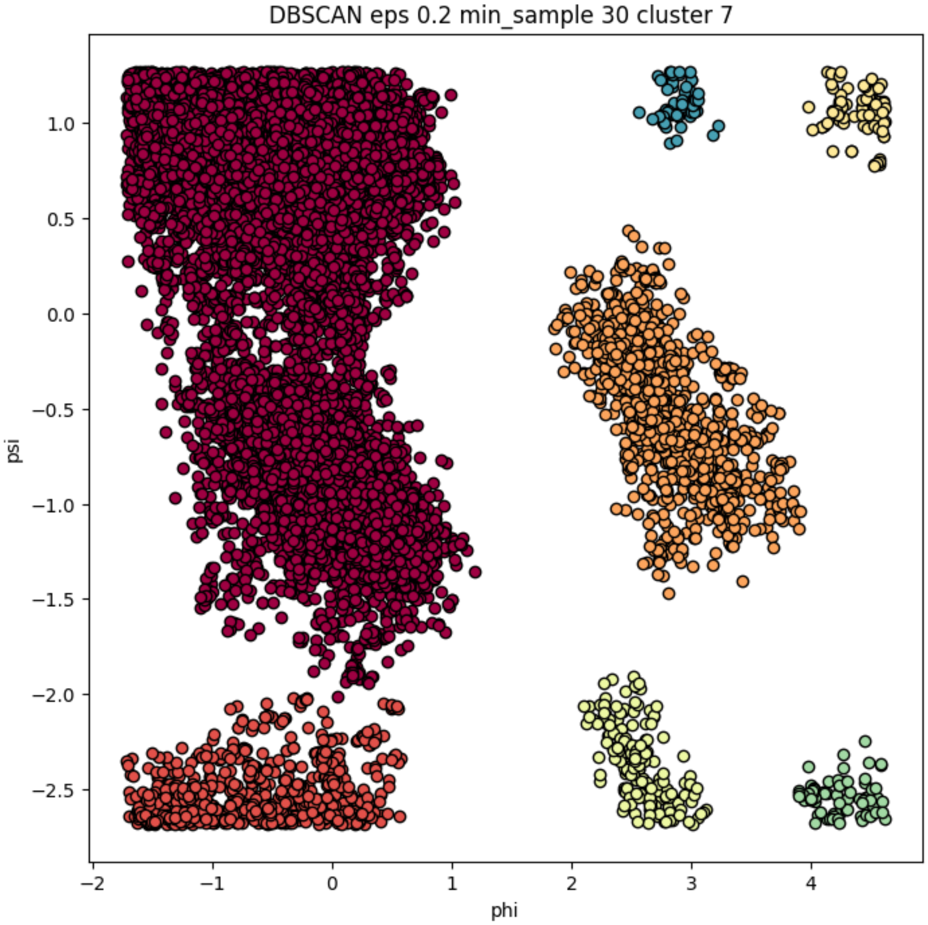

Figure 5: DBSCAN with 7 clusters and noise removed

3.a Motivate the choice of: i. the minimum number of samples in the neighbourhood for a point to be considered as a core point, and ii. the maximum distance between two samples belonging to the same neighbourhood ("eps" or "epsilon"). Compare the clusters found by DBSCAN with those found using K-means.

i. When selecting a reasonable minimum number of samples you have to consider which clusters are meaningful. Because if you set it too high, the algorithm may miss some clusters that are meaningful and if you set too low, then the algorithm may identify many small clusters or noise that are not meaningful. By testing a range of samples you can determine what clusters are meaningful. In this case a reasonable number of samples was 30, this sample number creates 7 distinct clusters and removes all necessary noise.

ii. The maximum distance between two samples (epsilon or eps) determines a radius around each point and is a crucial parameter when separating clusters from noise. Points within this radius are considered as part of the same cluster, while points outside are seen as separate clusters or noise. Since there is no one-size-fits all value for epsilon, and it must be chosen based on the data, we decided to test a range of values to find what fits our data the best. In this case a reasonable value for epsilon was 0.2. This value was acceptable because any higher values created large incoherent clusters, while any lower values seemed to heavily distort the ”structure” of the data and remove meaningful clusters.

When comparing the results found by DBSCAN and those found using K-means there are a few important differences. Firstly the number of clusters are different, this makes sense because the amount of clusters found using K-means is decided by us and DBSCAN determines the amount of clusters automatically. From the clusters identified we can clearly see that K-means does not perform well with outliers. DBSCAN, on the other hand, is significantly more robust when it comes to outliers and will identify them as separate clusters or noise.

3.b Highlight the clusters found using DBSCAN and any outliers in a scatter plot.

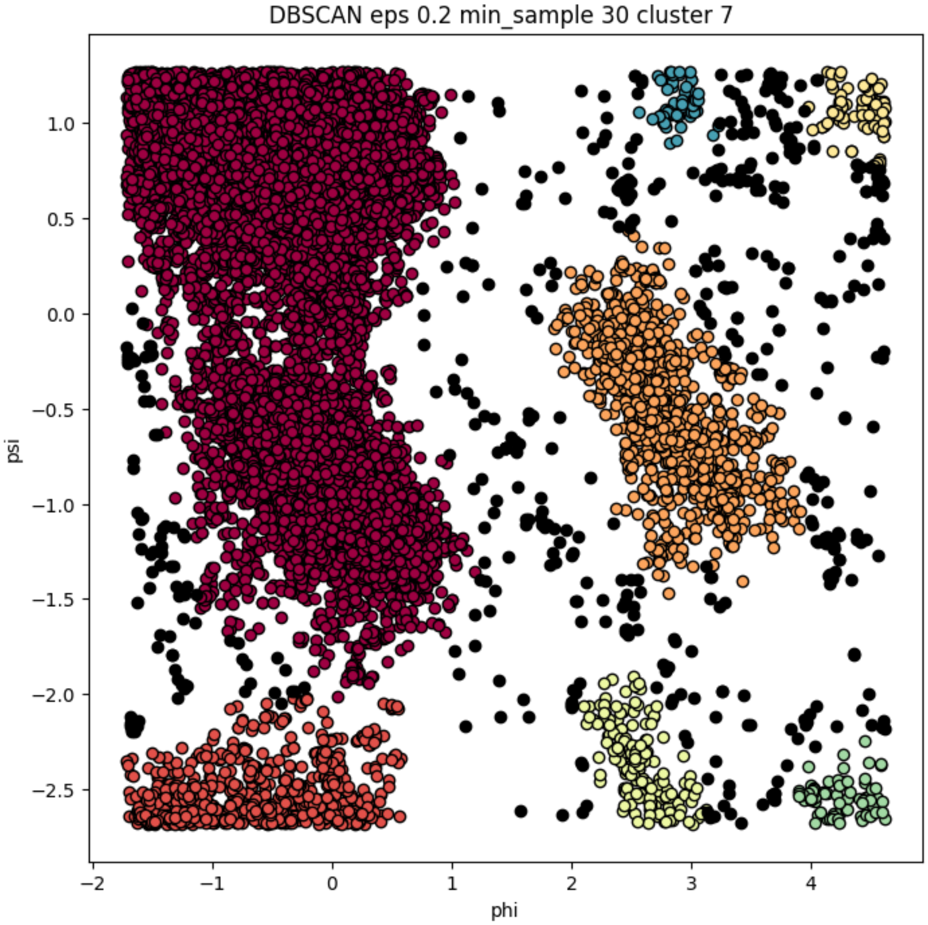

Figure 6: DBSCAN with the outliers

The outliers, which are the points that don’t belong to any cluster, are plotted as black dots. The estimated number of outliers are around 520 points.

3.c How many outliers are found? Plot a bar chart to show how often each of the amino acid residue types are outliers.

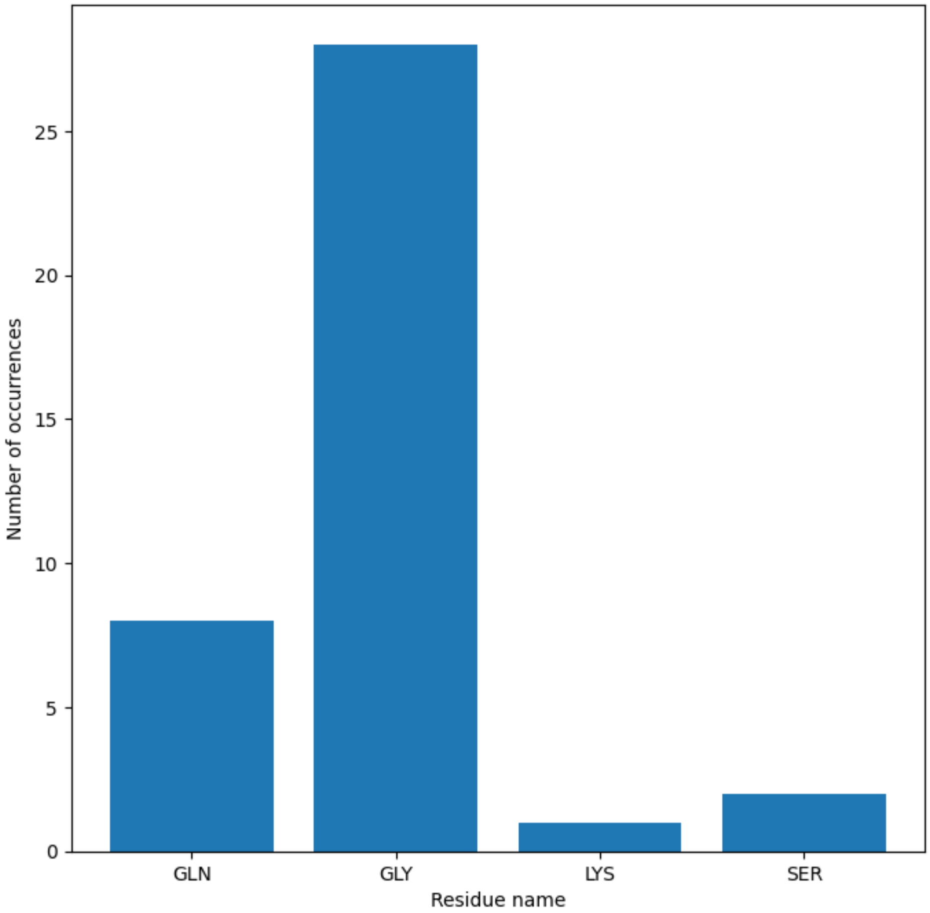

Figure 7: Bar-chart of most common amino acid residue outliers.

By making a bar chart, we notice that the most frequent amino acid outlier is GLY, in second place GLN and thereafter SER and LYS.

4 The data file can be stratified by amino acid residue type. Use DBSCAN to cluster the data that have residue type PRO. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters (i.e., the clusters that you get from DBSCAN with mixed residue types in question 3). Note: the parameters might have to be adjusted from those used in question 3.

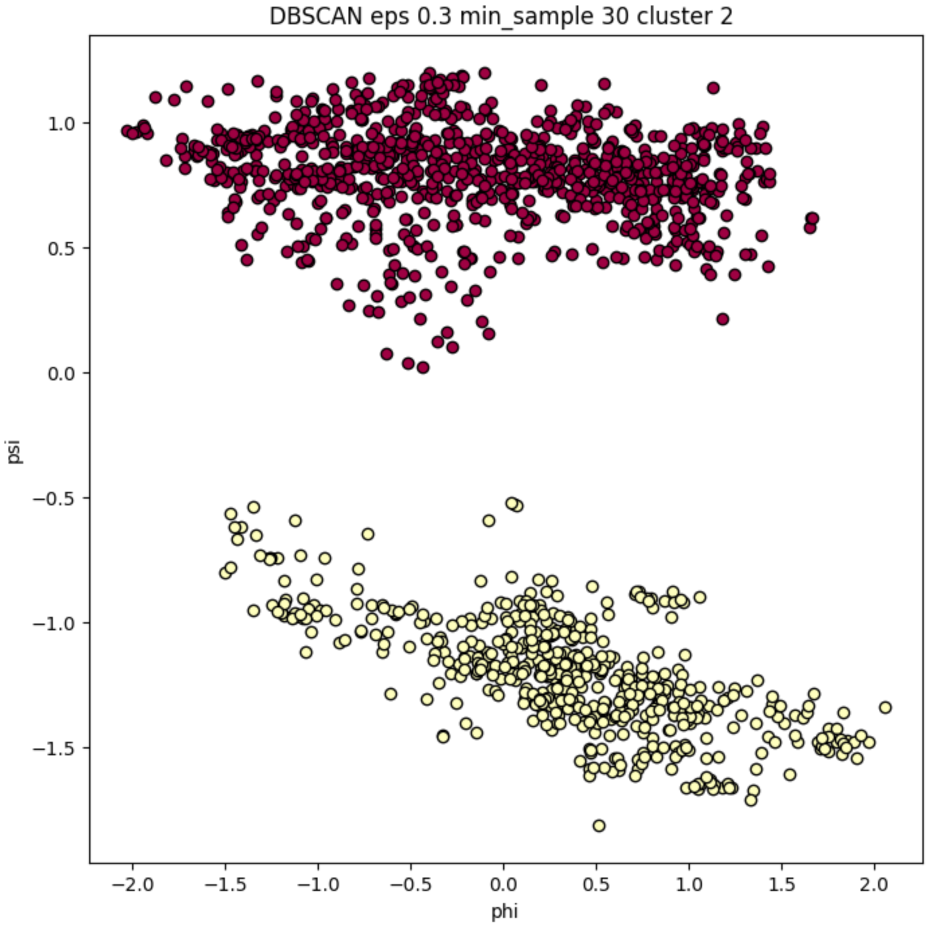

Figure 8: DBSCAN on amino acids with residue type ”PRO”.

When doing the DBSCAN with the ”PRO” residue we get the figure8 which has 2 clusters. These two clusters seems to be related to the big red cluster in the mixed type DBSCAN. Since there are only 1596 ”PRO” amino acids the overall structure and density of the mixed typed DBSCAN is not effected by the ”PRO” type amino acids.

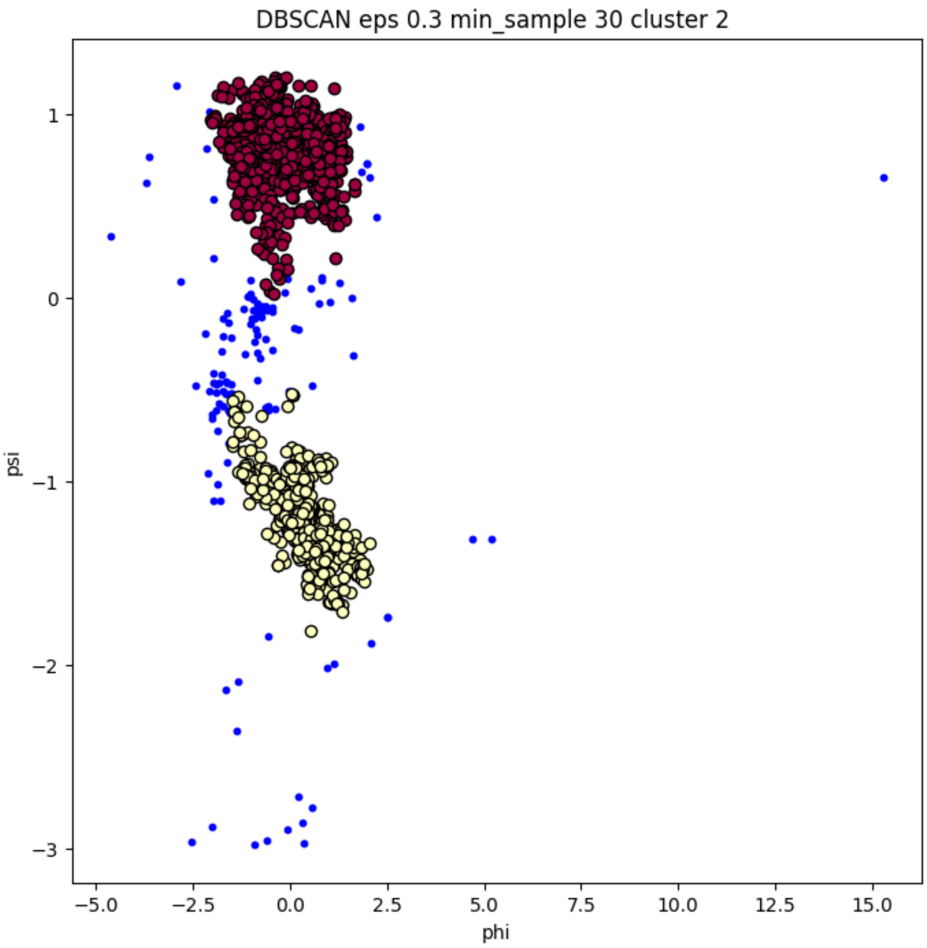

Figure 9: A scatter plot with amino acid of type ”PRO” with a DBSCAN layered on top

In figure 9 we can how the two clusters would roughly look in the mixed type DBSCAN.