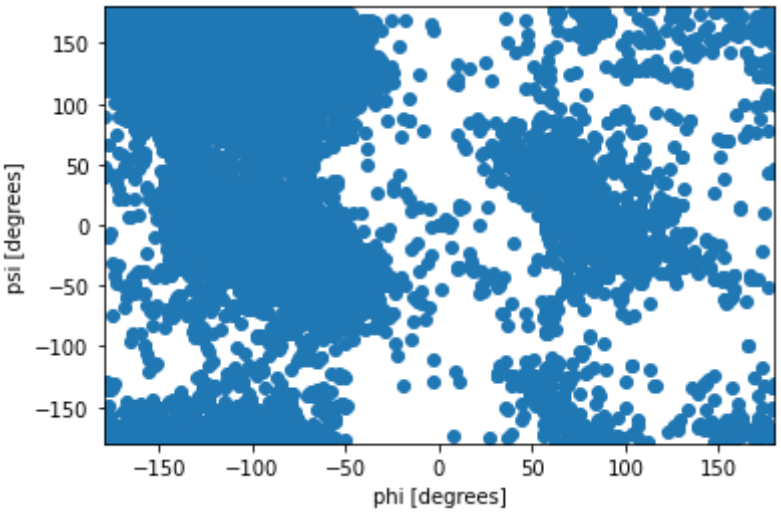
Assignment 3 - Clusters

Group 62
Fredrik Lilliecreutz & Erling Hjermstad

Time spent on the module:
Fredrik Lilliecreutz: 10 hours
Erling Hjermstad: 10 hours

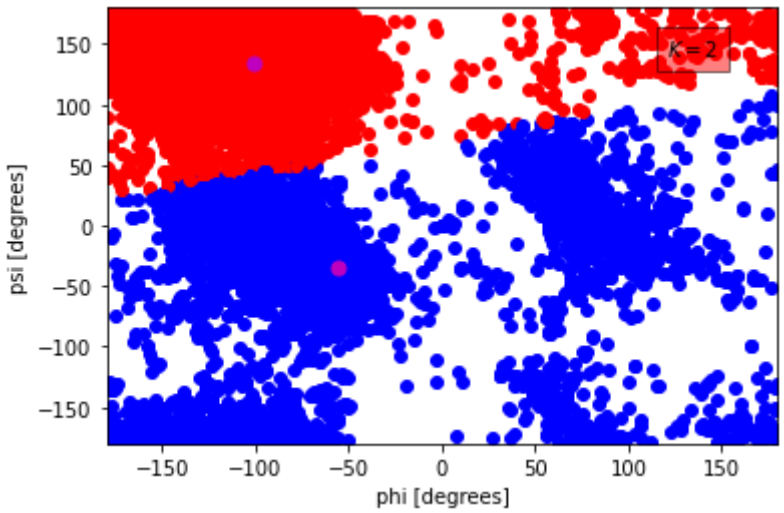
1.
Draw a scatter plot that shows the phi and psi combinations in the data file.

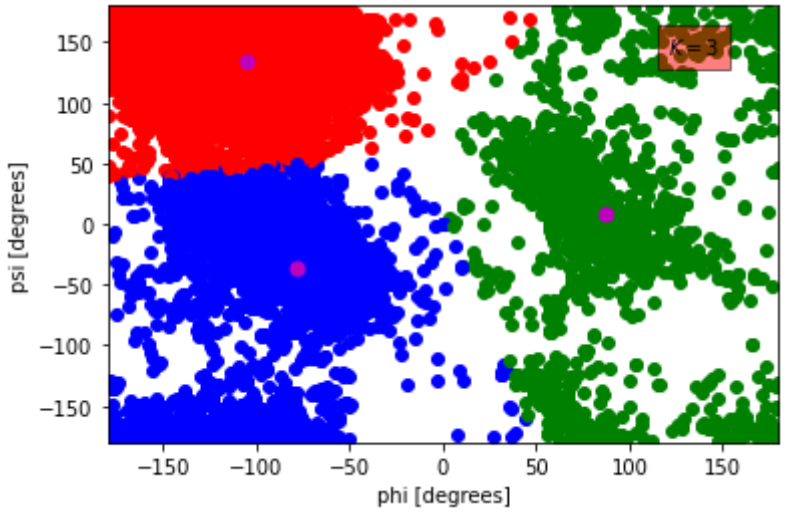


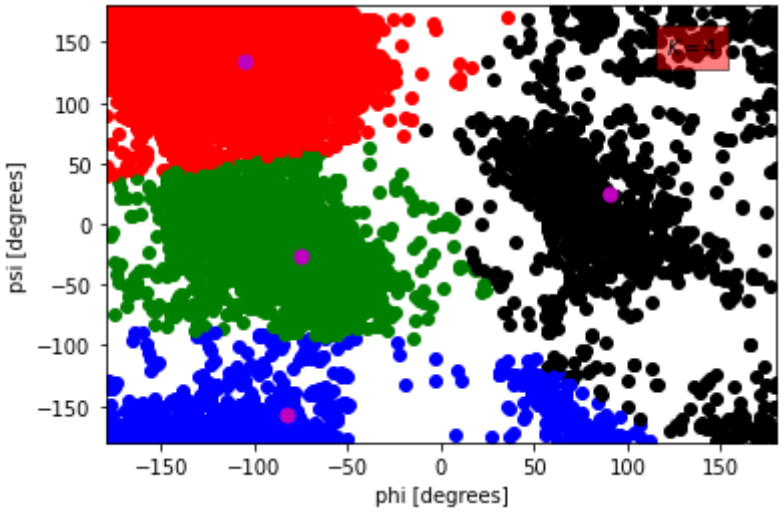
Axis are set to [-180, 180] since this is a rotation, meaning -180 and 180 actually are the same angle.

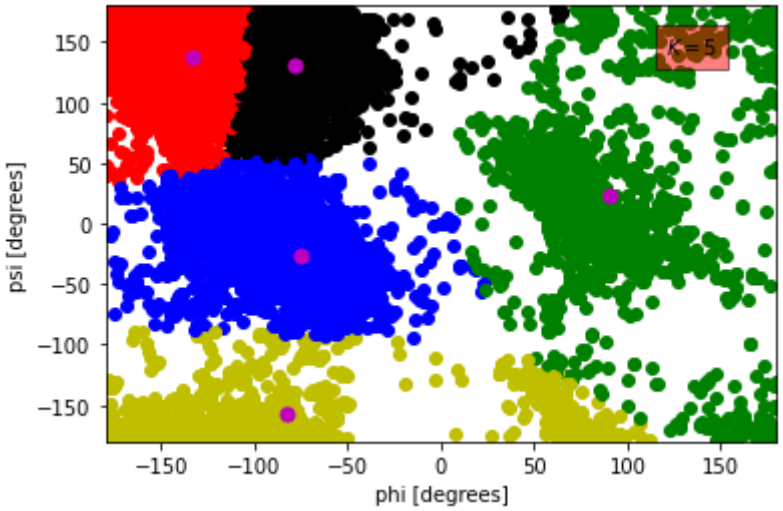
2.
Use the K-means clustering method to cluster the phi and psi angle combinations in the data file.

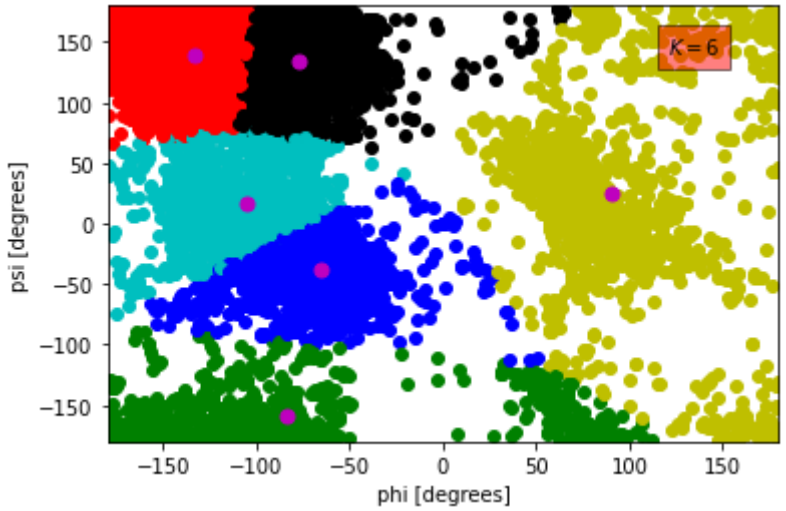
a.
Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.

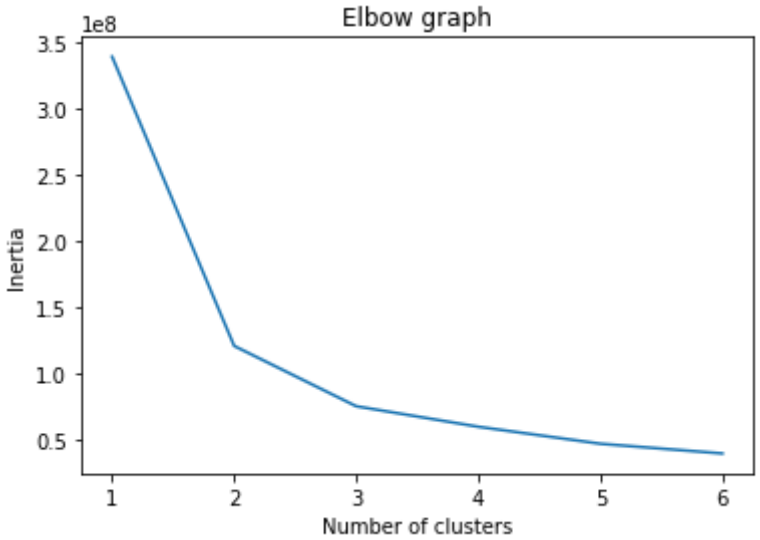

K = 2
psi [degrees]
phi [degrees]
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150


K = 3
psi [degrees]
phi [degrees]
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150


K = 4
psi [degrees]
phi [degrees]
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150







Visually K = 3 or 4 seem like the best choices, but when calculating the total inertia for the clusters K = 2 or K = 3 stand out as the moment when the effect of adding more clusters decreases significantly. Still, we would argue that K = 4 is a decent choice because the small 4th group in the left bottom corner does stand out visually. However, a different separation (left-right) with K = 2 also looks fitting and might be even better as we have no guarantee of

optimal convergence with k-means. We would suggest K = 3 as the best choice based upon the elbow graph above. This was chosen because the graphs flatten at k =3.

b.

Validate the clusters that are found with the chosen value of K.

The question was withdrawn by the teacher.

c.

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

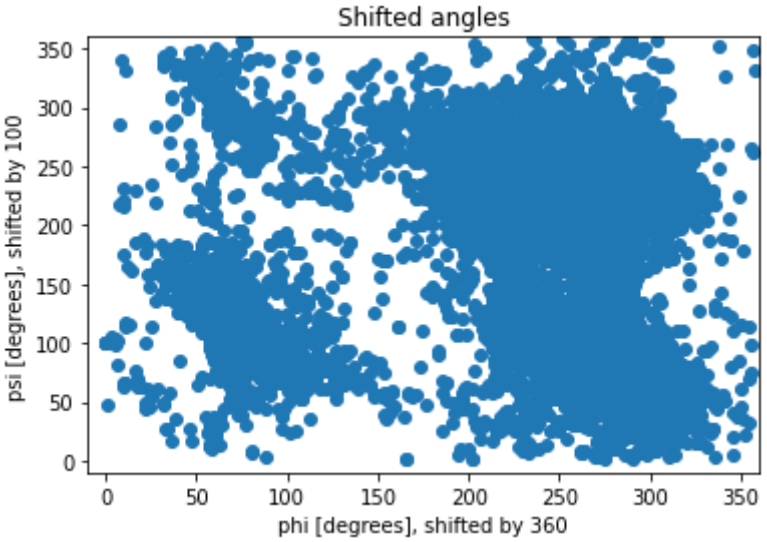
There is some uncertainty with the visualization versus the elbow graph when it comes to the clusters. However, in d) we have solved how to visualize the data better and then we could tell that the optimal amount of clusters is 3, this does stand out the clearest when using a subset of the entire dataset. The smallest group is in fact a part of the biggest group, but the angle has wrapped around from 180 to -180 degrees. This made the results very reasonable, and it looks like 3 is obviously the right value for K.

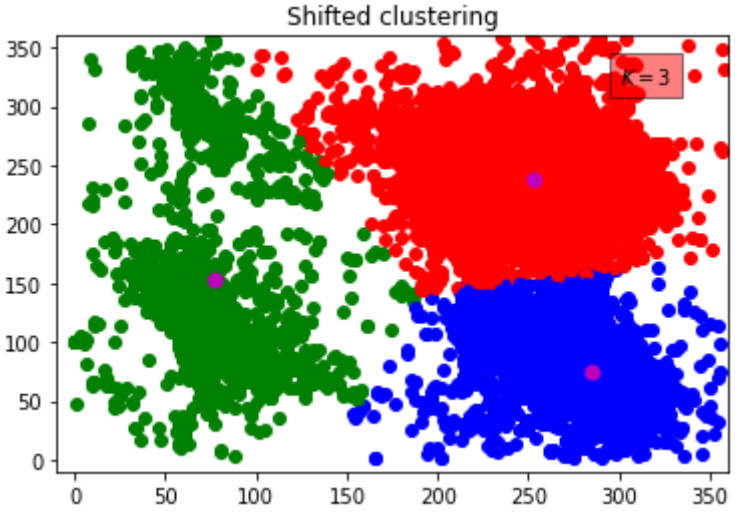
d.

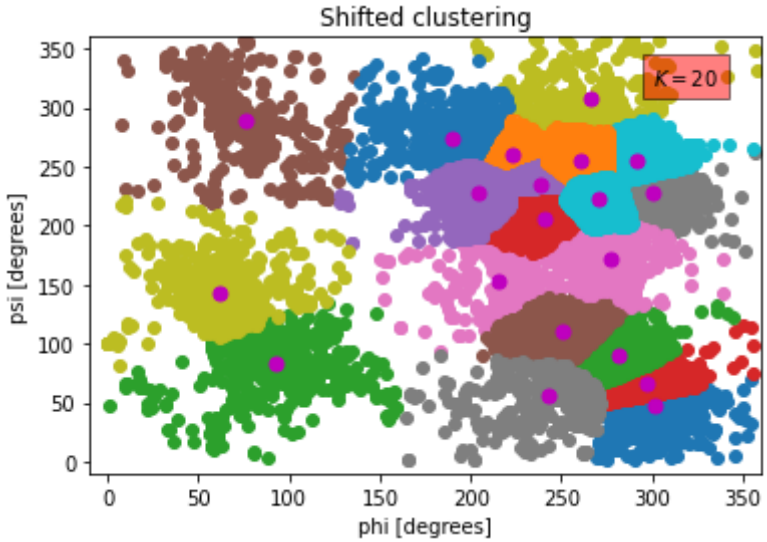
(For A Higher Grade) Can you change the data such that you reach better results? (Hint: since both phi and psi are periodic attributes, you can think of shifting them by some value and then use the modulo operation.)

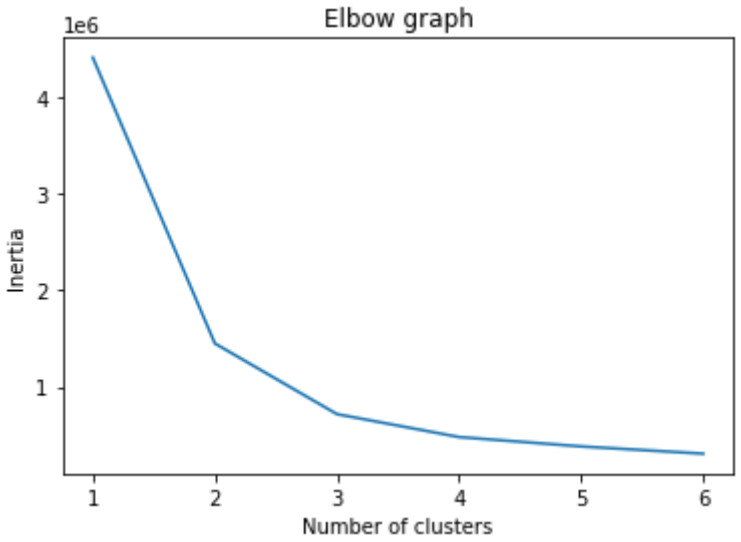
After examining the clusters further, we realized that the data points around <=150 and >=150 are very close to each other but misrepresented in the first graph. Because rotating nearly 180 degrees in both directions will end up very close to each other. This explains why there was a very small cluster close to -180 degrees psi. When looking at the result now the clusters look reasonable.

We solved the task by shifting the psi with 100 degrees and shifting phi with 360 degrees and doing modulo 360. This corresponds to changing where the angle is measured from. Now there were fewer points close to the edges of the graph. Phi and Psi are now in the [0, 360) set.


Shifted angles
psi [degrees], shifted by 100
phi [degrees], shifted by 360
0
50
100
150
200
250
300
350


Shifted clustering
K = 3
0
50
100
150
200
250
300
350


Shifted clustering
K = 20
psi [degrees]
phi [degrees]
0
50
100
150
200
250
300
350



Going forward, the shifted dataset is used.
There exist 20 types of amino acids, but 20 different classifications does not give a logical grouping. Visually it is hard to determine from phi and psi angles what amino acid a dot is.

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

a.
Motivate:
i. the choice of the minimum number of samples in the neighborhood for a point to be considered as a core point, and

Since we had a data set with nearly 40 000 data points we realized that we needed to have a quite big minimum number of samples, so we tried different numbers ranging from 20 - 100. A minimum number of 100 gave us the best result and it seems to be reasonable to us. So the method of finding the right minimum number was basically just by trying different numbers and making adjustments. Worth mentioning is that we first started out with the data set with just 500 data points and it was harder to find suitable clusters together with a reasonable eps.

ii. the choice of the maximum distance between two samples belonging to the same neighborhood (“eps” or “epsilon”).

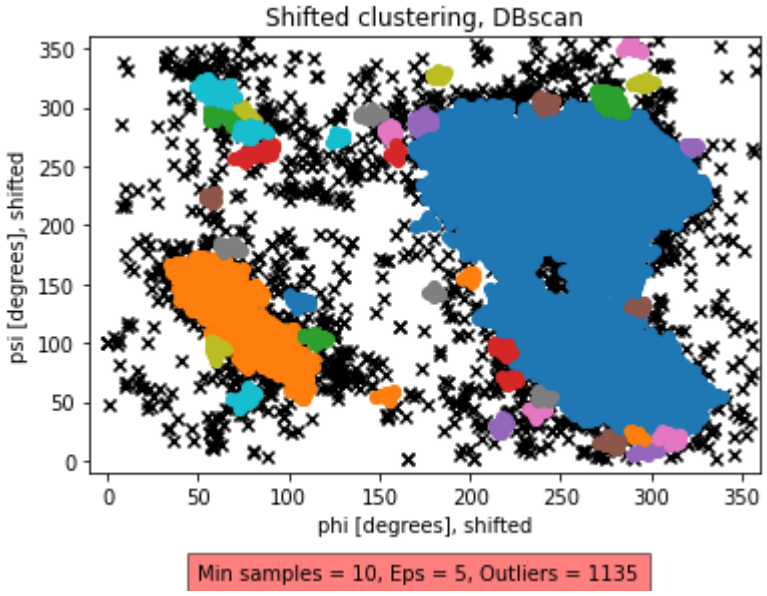
We tried different values and we discovered that this highly depends on the size of the dataset. We were using the dataset with 500 data points and to optimize the eps we would have to use a larger dataset. However, in our case, we would have to use a larger eps than we would if we had a bigger set. This is because there are less data points in every cluster(lower density) and we have to catch the data using a larger radius. However, a larger data set would give us the ability to catch noise better because we would get better clarity of the noise that is close to the cluster. It is easy to visually see noise, in the form of extreme

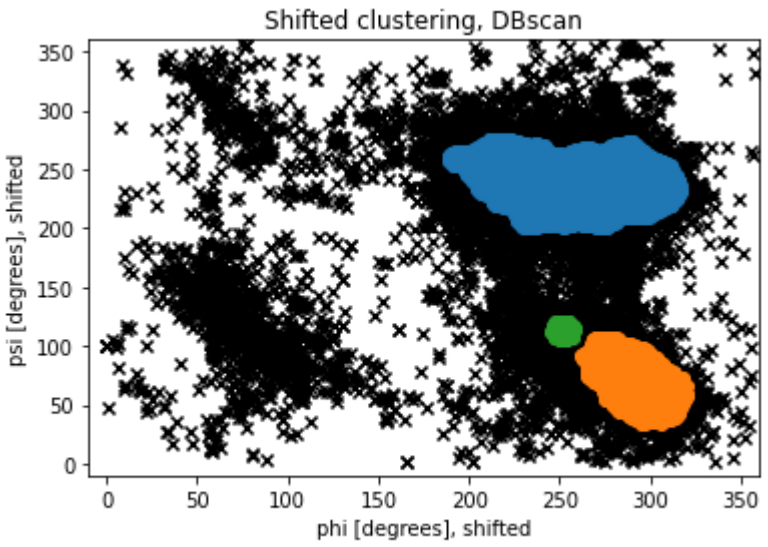
values in the graph but it gets harder to catch noise that is close to the cluster. And using a bigger radius and smaller minpts does not give as accurate results as we would like to have. To get the most distinct clusters, the optimal way of finding this would be to have as low epsilon as possible as well as the highest minpts. This will give us very distinct clusters but most of the data would be classified as noise.

After using the dataset with 500 data points, we realized that we had to use a larger dataset. This gave us a distinctively better result than using 500 data points, and we decided to use an epsilon of 10 which seemed reasonable to us. This was selected by trying different sizes and therefore selected the most reasonable epsilon size. When we chose too small epsilon we got too many clusters and too big resulted in one massive cluster. It is however important to understand the data behind the model before deciding the optimal epsilon since the epsilon decides how similar the data must be to belong in the same cluster.

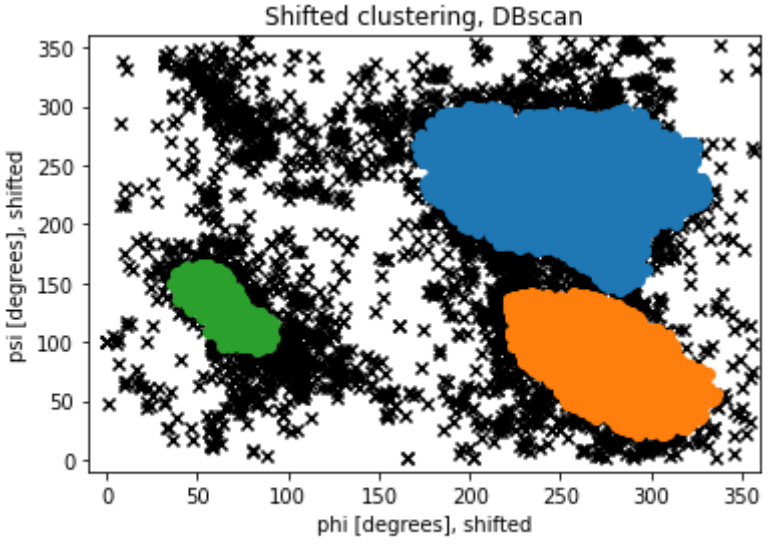
b.

Highlight the clusters found using DBSCAN and any outliers in a scatter plot. How many outliers are found? Plot a bar chart to show which amino acid residue types are most frequently outliers.

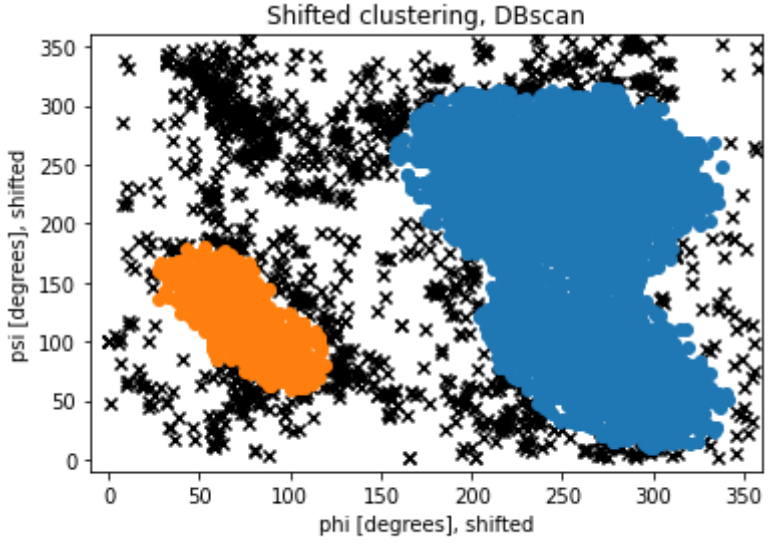




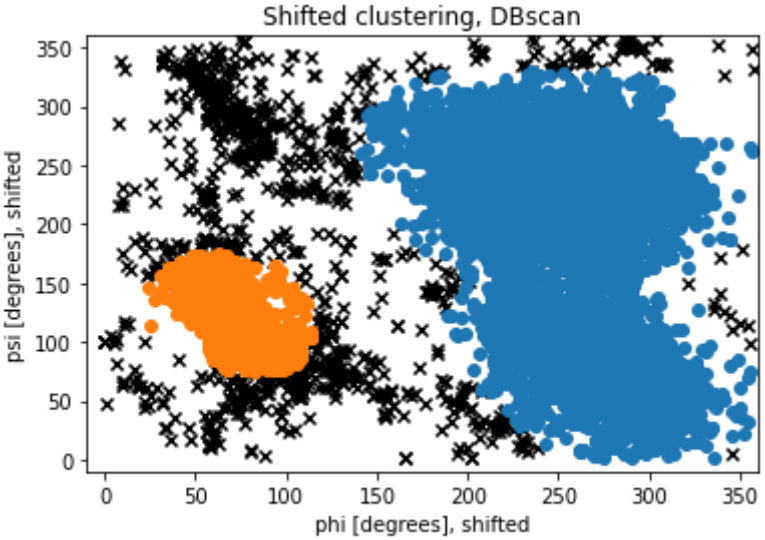
Min samples = 100, Eps = 5, Outliers = 8841



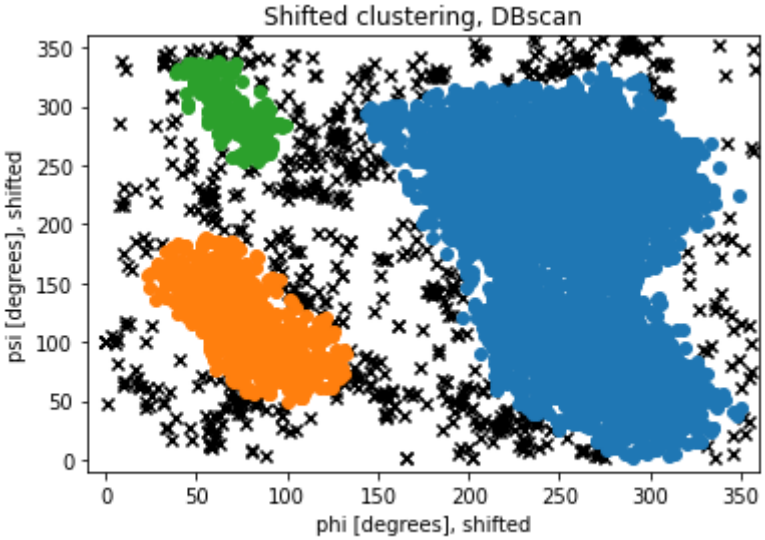
Min samples = 100, Eps = 10, Outliers = 2604



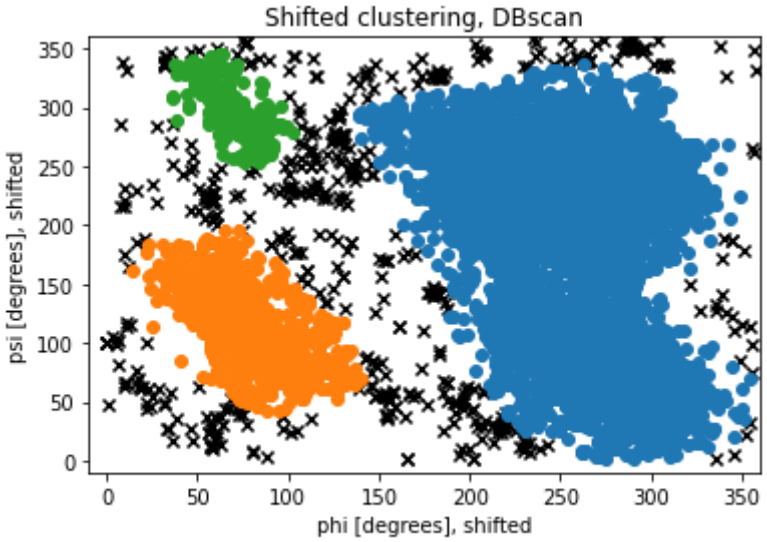
Min samples = 100, Eps = 15, Outliers = 1288



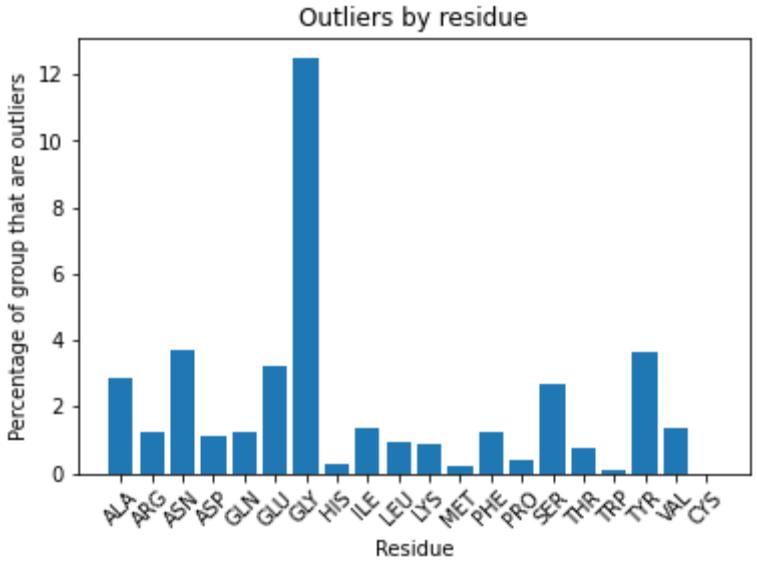
Min samples = 500, Eps = 30, Outliers = 1035



Min samples = 50, Eps = 15, Outliers = 723



Min samples = 80, Eps = 20, Outliers = 532



For all values GLY is by far the biggest contributor to the outliers, without being any more represented than the others.

c.

Compare the clusters found by DBSCAN with those found using K-means.

When comparing the clusters found by DBSCAN with K-means we noticed that the DBScan finds more distinct clusters, i.e, more compacted clusters where most of the data is. In contrast to K-cluster it also has a function to detect noise which contributes to these distinct patterns. In K-means we saw that some data points were included in a cluster whilst it was classified as noise in the DBSCAN method. This was clearly a difference between the two methods. If we draw a parallel to the previous lab when we were looking at house prices in Landvetter, the DBcluster would be more precise with the valuations in specific geographical areas.

To conclude the difference, DBSCAN finds noise that K-clustering does not, which makes the clusters more distinct.

d.

Discuss whether the clusters found using DBSCAN are robust to small changes in the minimum number of samples in the neighborhood for a point to be considered as a core point, and/or the choice of the maximum distance between two samples belonging to the same neighborhood (“eps” or “epsilon”).

The clusters found using DBScan are robust regarding the minimum number of samples, however, it is very sensitive when it comes to the eps. But this really depends on what small changes are defined as. But we could tell that the data is more sensitive to changes in epsilon than the minpoints. The radius tells us how similar each core point must be to each other to be in the same cluster and when using a large dataset it is reasonable that the DBMethod will be sensitive to small changes in the epsilon. This is because it tells us the accuracy of the clusters. In our case, we don’t have much experience in biology and therefore it is hard to understand the importance of each degree in rotations of the molecules. So we are adjusting the epsilon so it fits the number of k’s indicated by the elbow method. When it comes to the minimum amount of points it is important to have a high enough number to avoid noise data. Since our data set was relatively big, we had to have a large number of minpoints to achieve the optimal result. Looking at absolute numbers, this was not as sensitive as the epsilon. Since the dataset is so big the minpoints parameter is less sensitive than epsilon, since there are 30 000 points and only 360 degrees.

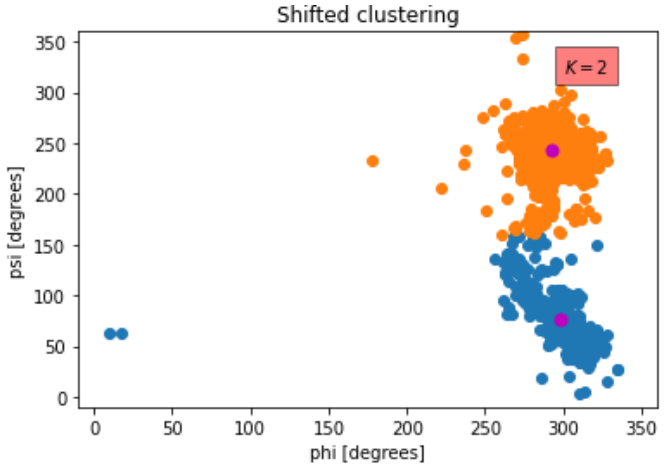
4.

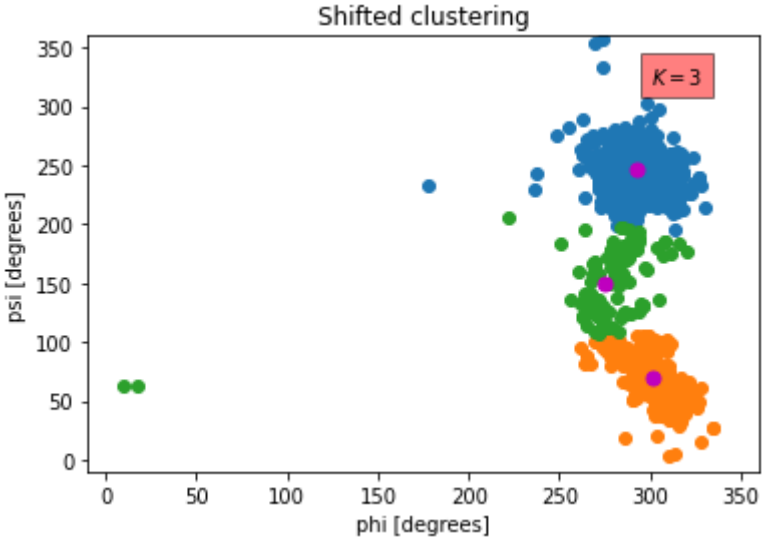
The data file can be stratified by amino acid residue type. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters. Similarly, investigate how the clusters found for amino acid residues of type GLY differ from the general clusters. Remember that parameters might have to be adjusted from those used in previous questions.

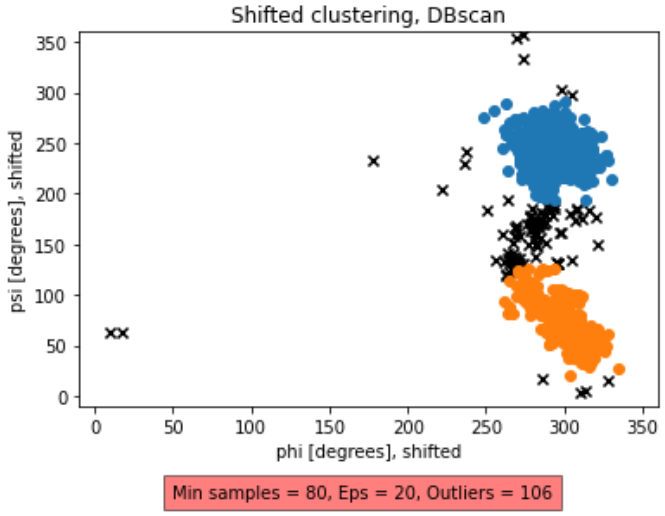
When looking at the plot of type PRO with the DBMethod we see that it looks like it is divided into two clusters. If we compare the K-cluster plot with the DB-method graph we remove the noise data which results in two distinct clusters. We also decide to narrow the epsilon to get even more distinct patterns. There is a group that is between the two clusters but with no distinct cluster. Even though this is a subset of the data set, we managed to keep the parameters relatively similar to the previous task. The density of the data is sufficient to keep these parameters and achieve the desired result.

The result of GLY was different, the data was spread out across the plot and visually it looked like there were 3-4 clusters. We tried different eps and minpts for this plot and we could conclude that this data, especially the top right, was spread out in different clusters when we reduced the eps. After trying a couple of combinations of eps and minpts we concluded that eps 15 and minpts 30 gave the best result for the data.
The fact that GLY is spread all over the plot gives some reason as to why it is overrepresented among the outliers.

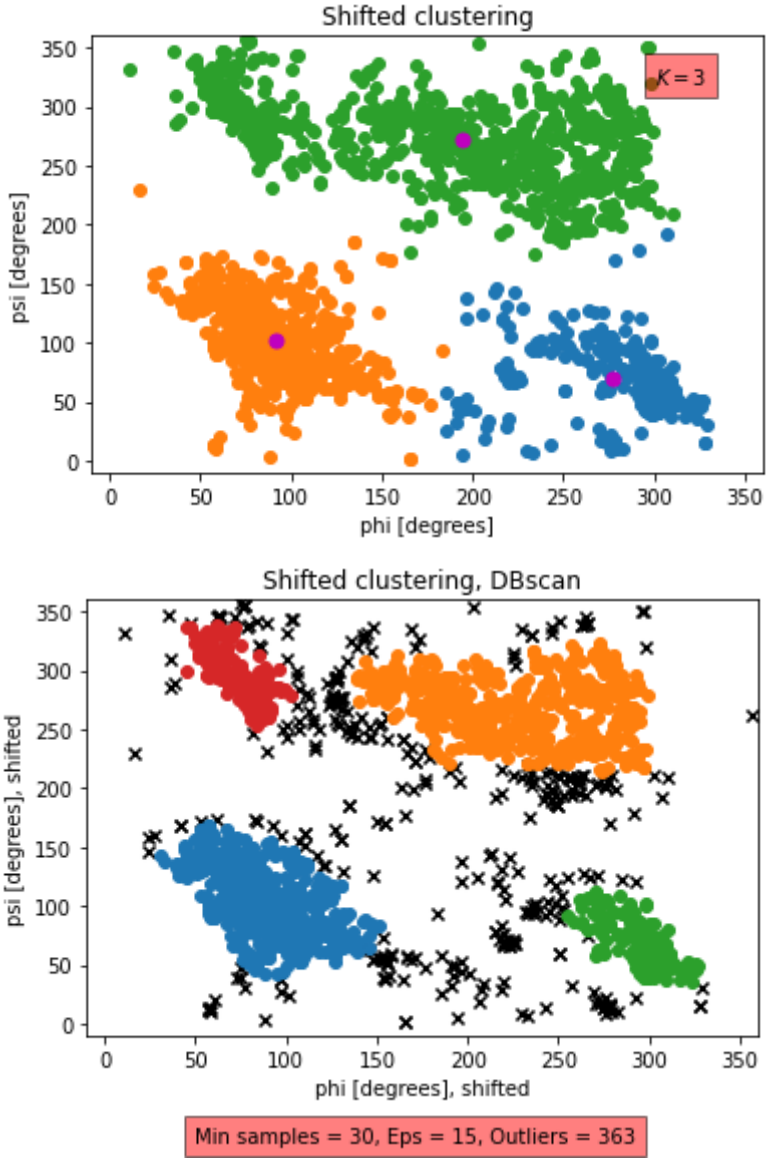
PRO:


Shifted clustering
K = 2
psi [degrees]
phi [degrees]


Shifted clustering
K = 3
psi [degrees]
phi [degrees]


Shifted clustering, DBscan
psi [degrees], shifted
phi [degrees], shifted
Min samples = 80, Eps = 20, Outliers = 106

GLY:


Shifted clustering
K = 3
psi [degrees]
phi [degrees]
Shifted clustering, DBscan
psi [degrees], shifted
phi [degrees], shifted
Min samples = 30, Eps = 15, Outliers = 363