Assignment 3 - Clustering

Nathan HAUDOT (4 hours), Hugo MATH (6 hours)

November 23, 2021

1 Distribution of phi & psi combinations

1.a Scatter plot

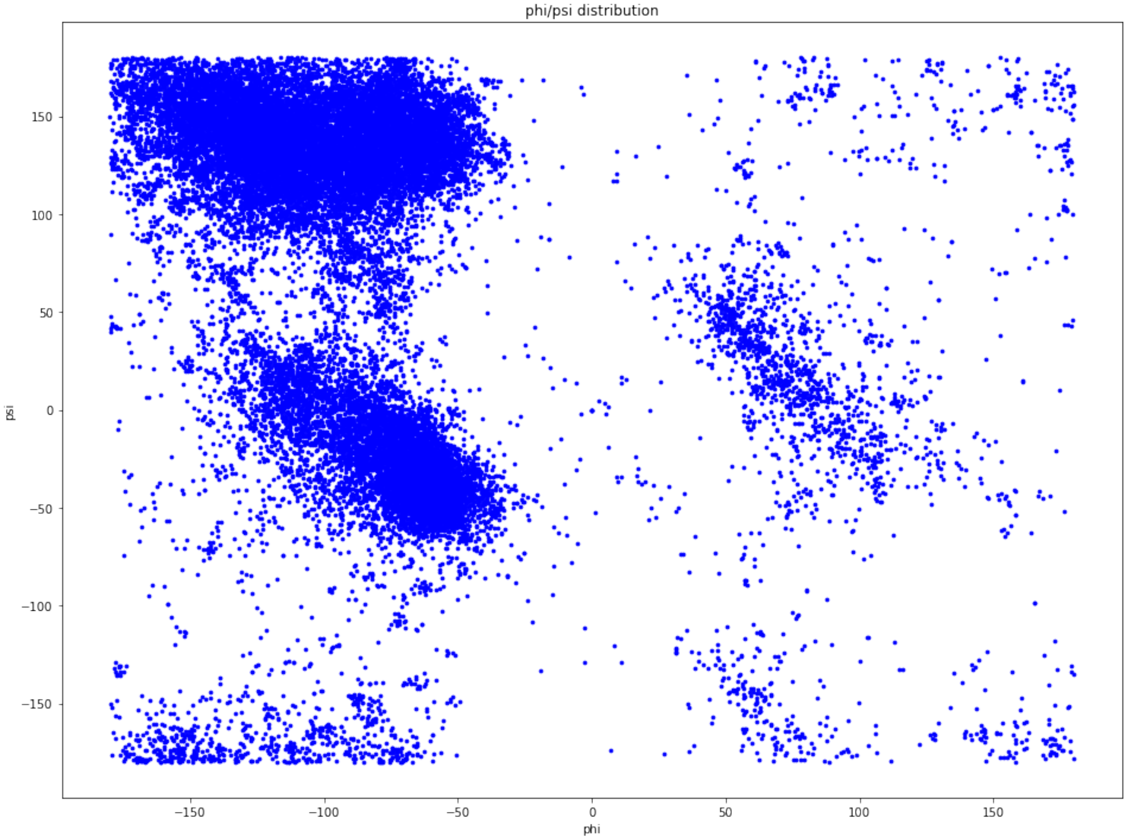

Figure 1: Distribution of phi vs. psi

1.b Heatmap

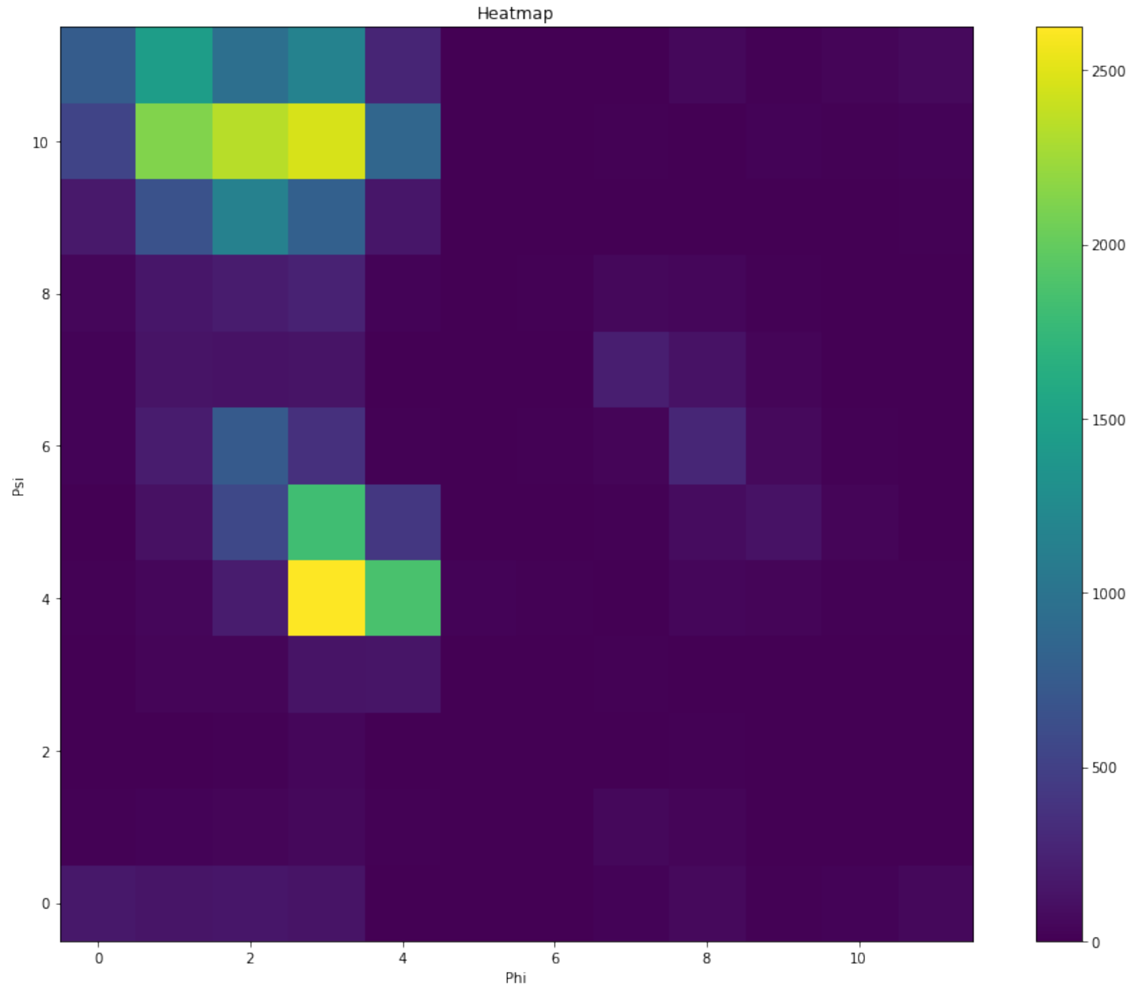

Figure 2: Heatmap of phi vs.psi binned at 12

As it is a heatmap, we are obliged to ”bin” our data. Here, we have chosen a binning of 12, because it delimits enough the supposed clusters we are trying to identify.

Like the scatter plot graph, we can already identify two or three clusters. This analysis will be useful later in the use of K-means clustering in order to choose a correct value for K.

2 K-means clustering method

2.a Eperimentation with K values

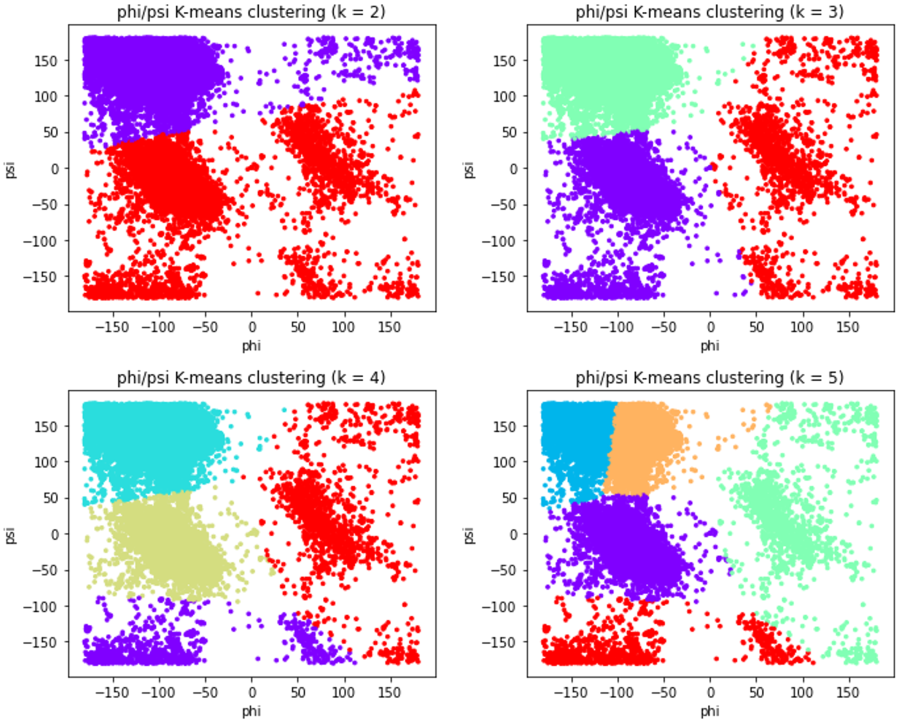

Figure 3: K-means clustering with multiple values of K

After experimenting with several values of K, we think that the appropriate value is 3. We already had this analysis using the heatmap in figure 2. If we increase the value of K (the number of clusters), the K-means method does not work as we would like, and splits the obvious clusters into multiple smaller clusters (see figure with k = 5).

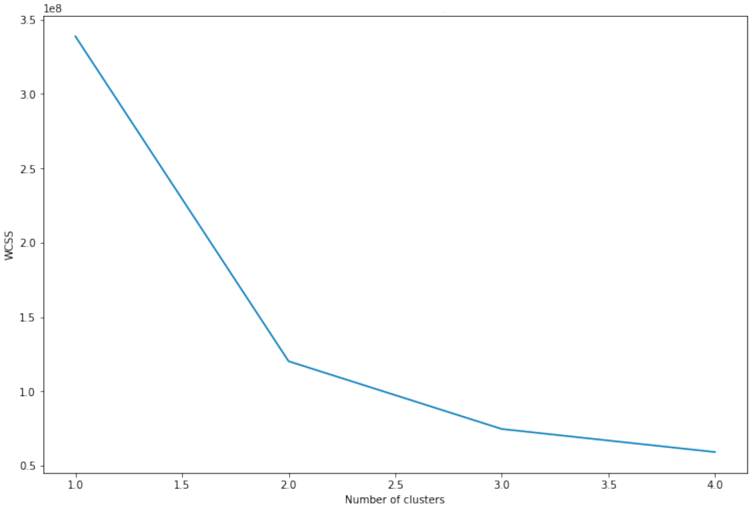

Figure 4: Elbow method for finding the best K value

We also used the elbow method to support our hypothesis, so our choice remains to be k = 3.

2.b Cluster validation with k = 3

To validate our choice of the number of clusters, we check the stability of the subsets by randomly removing a proportion of points from the dataset (10 thousand, 15 thousand, and 20 thousand points).

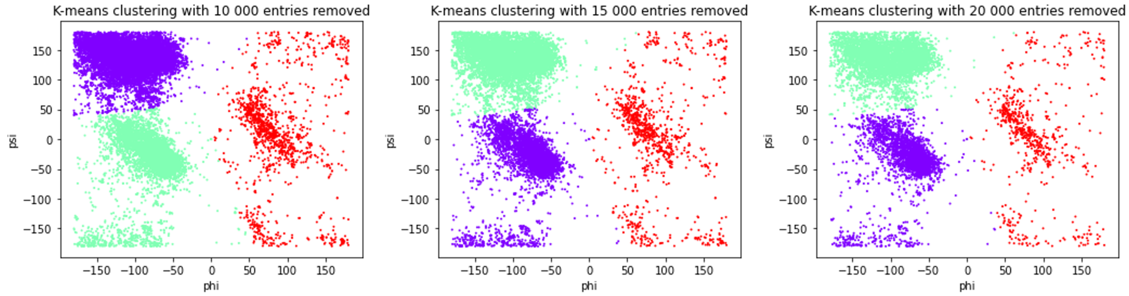

Figure 5: K-means clustering at k = 3 with removed entries

We see here that the clustering is always the same, only the cluster's color might change but there is no drastic change when removing points: the cluster stability is thus assured.

2.c Do the clusters found in part 2.a seem reasonable?

The clusters found in question 2.a. for k = 2 do not seem relevant to us, and beyond k = 4, we start to see clusters that we would never have guessed before. Since the heatmap graph, we think that we can intuitively identify 3 clusters. The results of the analyses carried out afterwards (K-values experimentation, elbow method, cluster validation) confirm once again this choice.

2.d Changing data to get better results

We know that our data dimensions are angles in degrees. Indeed by scaling them to get the radians and then by shifting by $\pi/2$, we can get clusters with much precise and clear decision boundaries. We can also use the **StandardScaler()** from *sklearn* which shift by the mean and scale by the standard deviation (we will use it with the DBSCAN method).

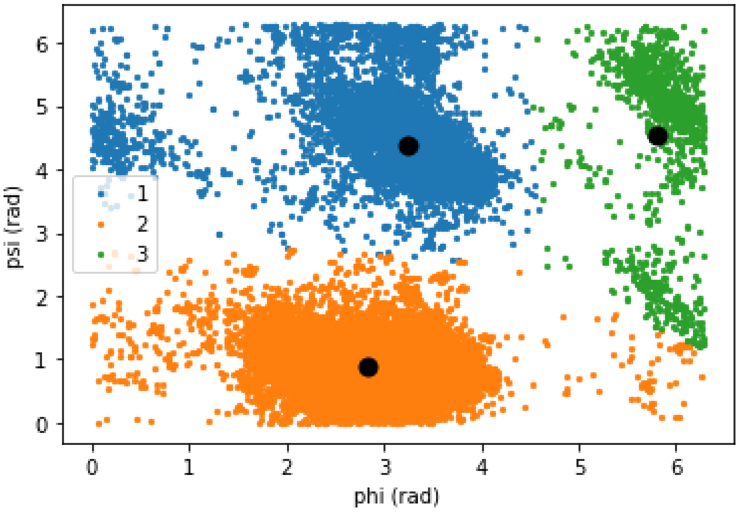

Figure 6: Shifted values cluster centroids

3 DBSCAN clustering method

Using DBSCAN, we can estimate the number of clusters and identify the noise by modulating two parameters : **epsilon** which is the maximum distance (euclidean distance), and the **minimum of samples** within this distance.

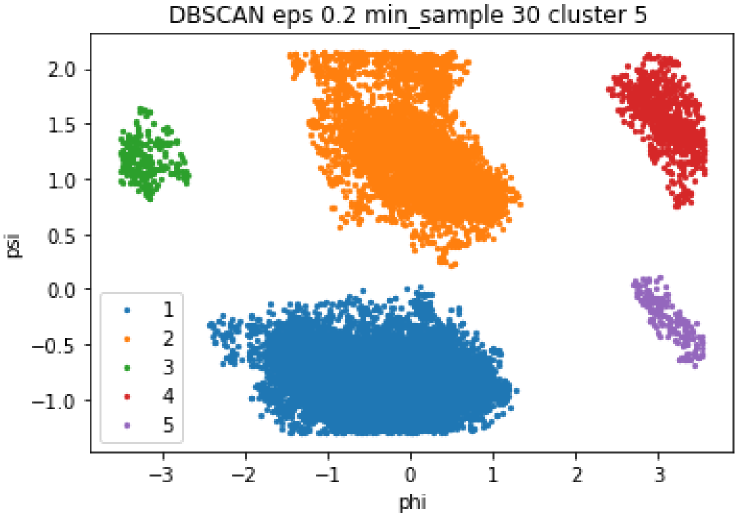

Figure 7: DBSCAN method with 5 clusters

3.a Choices motivations

i. The minimum of samples is directly related to the density. In fact DB stands for **Density Based**. Thus, we can see that our scatter plot is relatively crowded (we can see it also using the parameters of the *sklearn* DBSCAN model), so we have to have a relatively big min. of samples in order to suppress noise and group our clusters. We can start by putting an interval of $min_s = [20, 45]$.

ii. **Epsilon** refers to the radius in which we have our minimum of samples. This parameter is crucial to separate clusters from each other and from the noise. If we take a small distance, we will see hundred of small clusters. On the other hands, if we go for a very large epsilon, then we will get few clusters. In our case after shifting the data, it's more easy to separate the cluster by putting a medium to small **epsilon**. We can put an interval of $\epsilon = [0.1, 0.30]$.

The resulting parameters are $\epsilon = 0.2; min_s = 30$.

3.b Cluster finding highlights

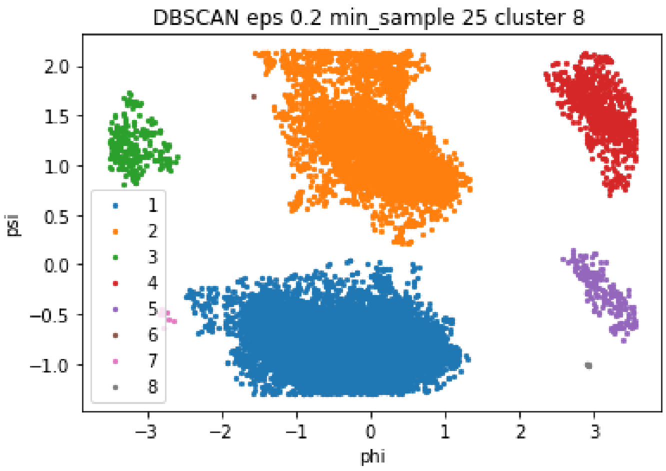

Figure 8: DBSCAN method with outlier clusters

Using a sample value of 25, we find 3 outlier clusters (6, 7 and 8) which are shown in the chart above.

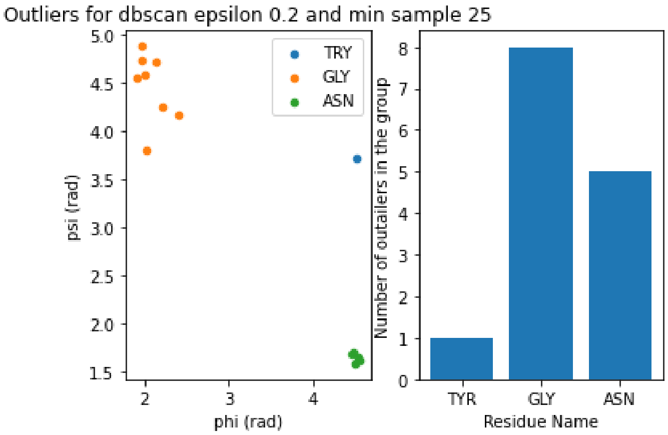

Figure 9: —

By making a bar chart, we notice that the most frequent amino acid outliers are GLY and ASN. There is only one TYR.

3.c Comparison between DBSCAN & K-means

We can state that K-means recognizes more often spherical clusters and it remains difficult to differentiate other forms of clusters (in the 2-dimensions). Moreover, DBSCAN, as we can see, is able to identify noise. This has a double-edge because by increasing too much the minimum sample, we can remove information out of the clusters.

3.d Is the clusters found using DBSCAN are robust to small changes ?

We conclude that one disadvantage of DBSCAN is its **sensitivity**. Indeed, it takes some time to find the good **hyper parameters**. We achieved this by heuristic which means incrementing our parameters with a step and going through an interval. Finally, K-means is more flexible in this way but may represent always spherical clusters whereas DBSCAN, if well tuned, can recognize more clusters types.

4 Amino acid comparaison

4.a a. PRO

After applying K-means and DBSCAN to our PRO residue dataframe, we got these 2 graphs:

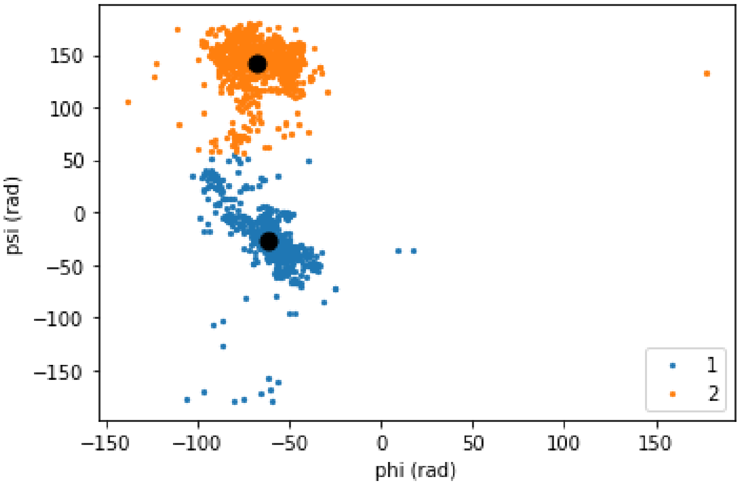

Figure 10: K-means for PRO residue with k = 2

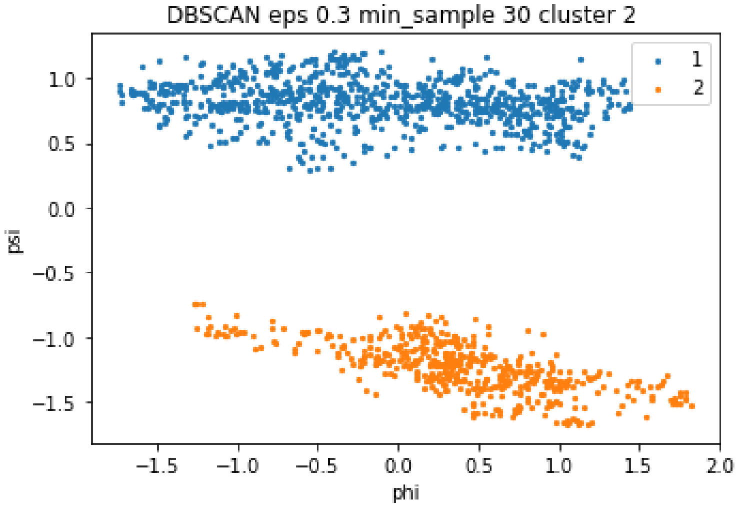

Figure 11: DBSCAN for PRO residue

With the PRO residue we only got 2 clean clusters using kmeans and DBSCAN. With all chains, we got 5 clusters and 3 with k-means. This indicates that all chains concatenate together to create multiple linear combination of the data. Adding the clusters all together creates another one but with different density. Indeed, there is 1597 rows for the PRO residue dataframe and 16 000 in total with all the residue, meaning that there is a low density in these two clusters and that finally it doesn't really impact our final cluster.

4.b GLY

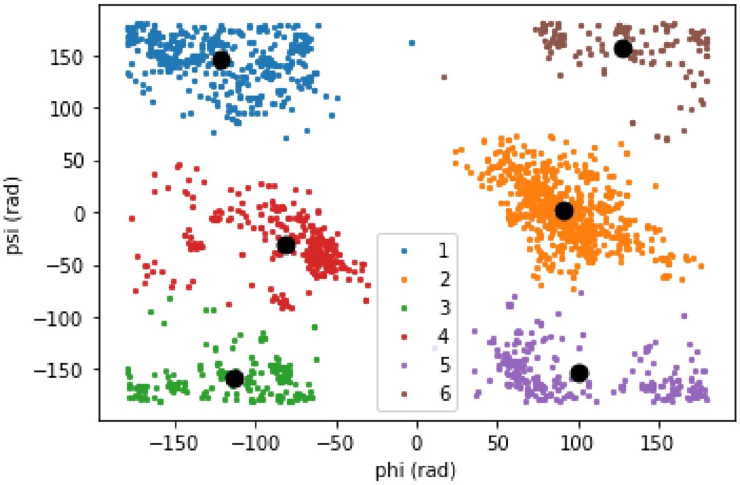

Figure 12: K-means for GLY residue with k = 6

We changed our parameters since we can clearly see 6 distinct clusters. This time, our clusters are less densed (that is why we decreased the minimum of sample to 20 for DBSCAN to keep the information). We can see here the big improvements with DBSCAN clustering method which is to remove a lot of noise near the clusters.

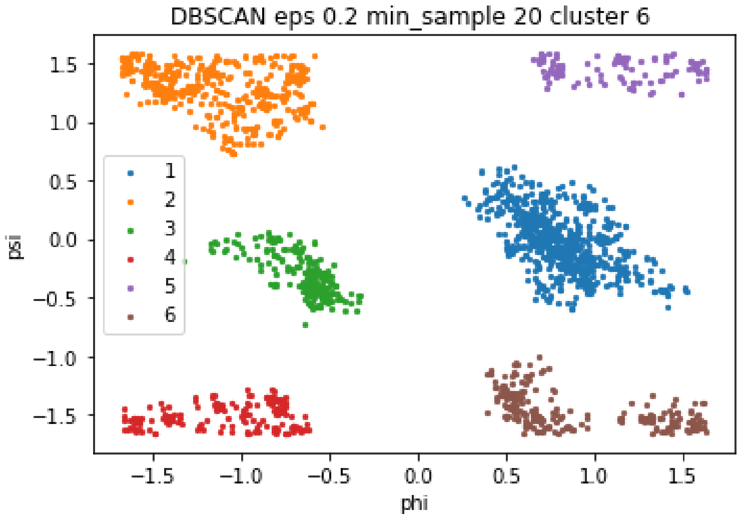

Figure 13: DBSCAN for GLY residue