Assignment 3

Kevin Jeryd, Vincent Hellner

Time spent: Kevin (4h), Vincent (4h)

November 22, 2021

Contents

1 Distribution of phi and psi combinations

Looking at combinations of phi and psi we made the following visualisations, see figures 1 & 2.

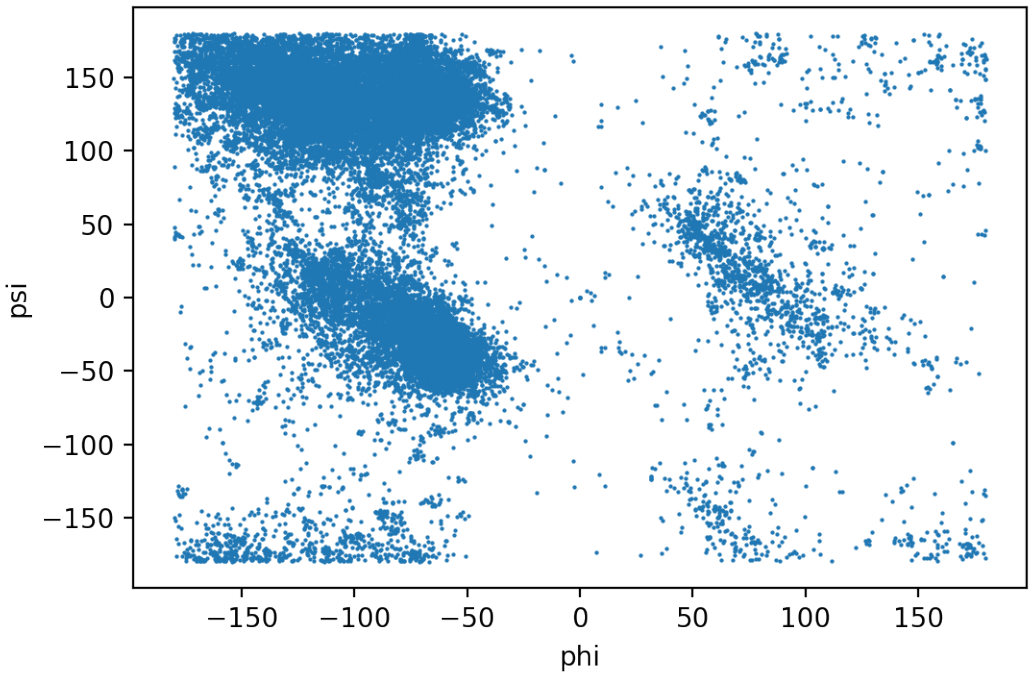

Figure 1: Scatterplot of phi and psi combinations

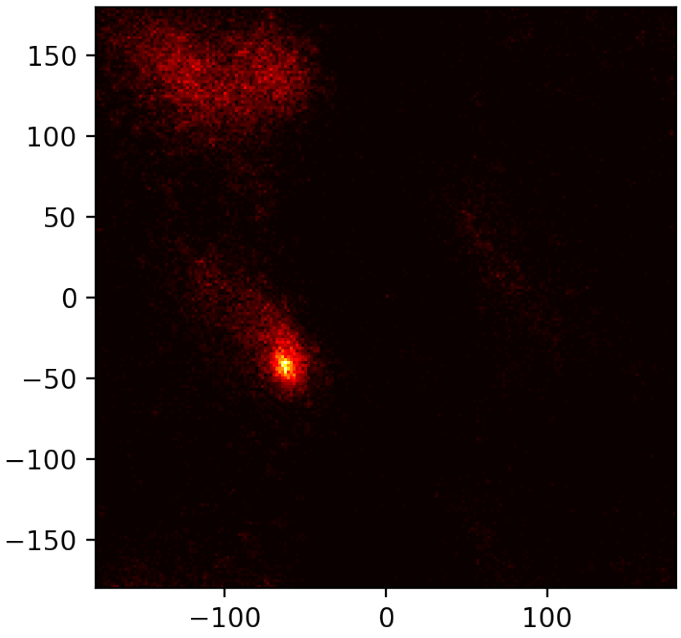

Figure 2: Heatmap of phi and psi combinations

2 K-Means Clusering of combinations

Using K-Means clustering to find groups in the data we started by using a K value of 3 as there are clearly at least 3 meaningful groups in the dataset, which are all clearly visible on the heatmap above.

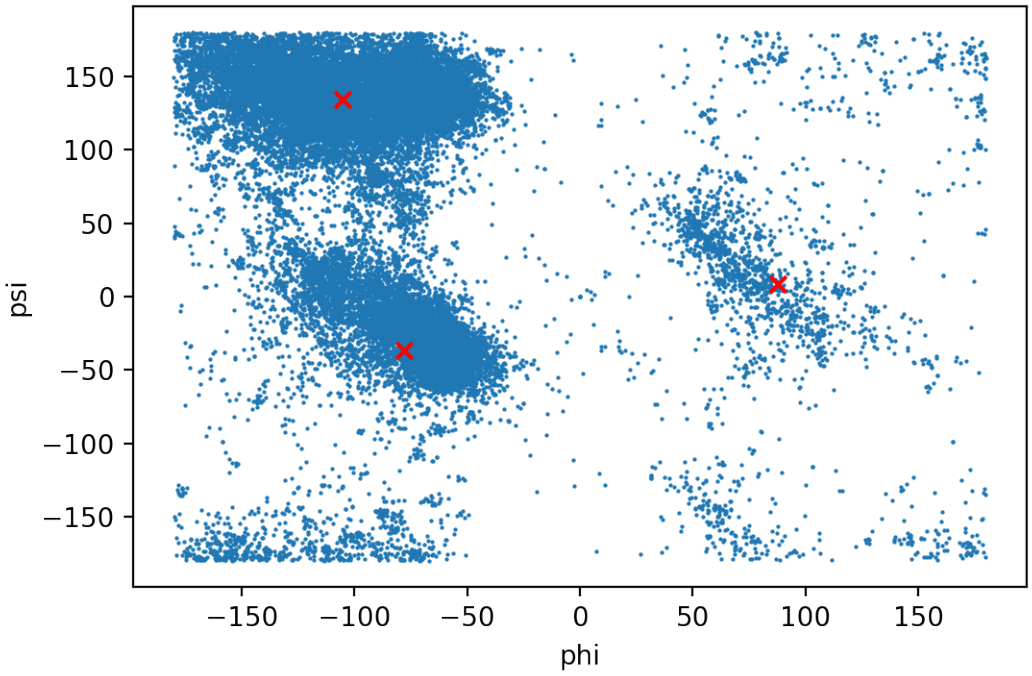

Figure 3: Cluster centers for K=3

Increasing the K value to 4 we can see that a new group has been found, although at further inspection we can see that the psi angle for this group is in the -180 area which means that it has simply just looped around and is just part of the group further up in the plot since the psi and phi attributes are periodic, see figure 4.

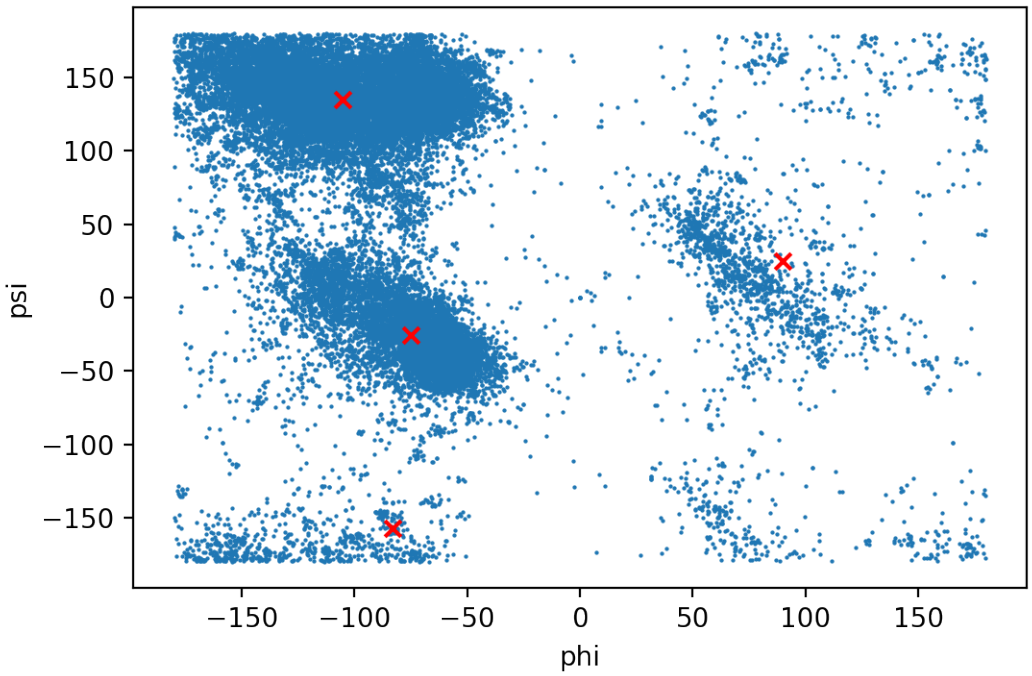

Figure 4: Cluster centers for K=4

This is even clearer when increasing the K value even higher (e.g 6) in which the algorithm simply creates more centers in the same groups, see figure 5.

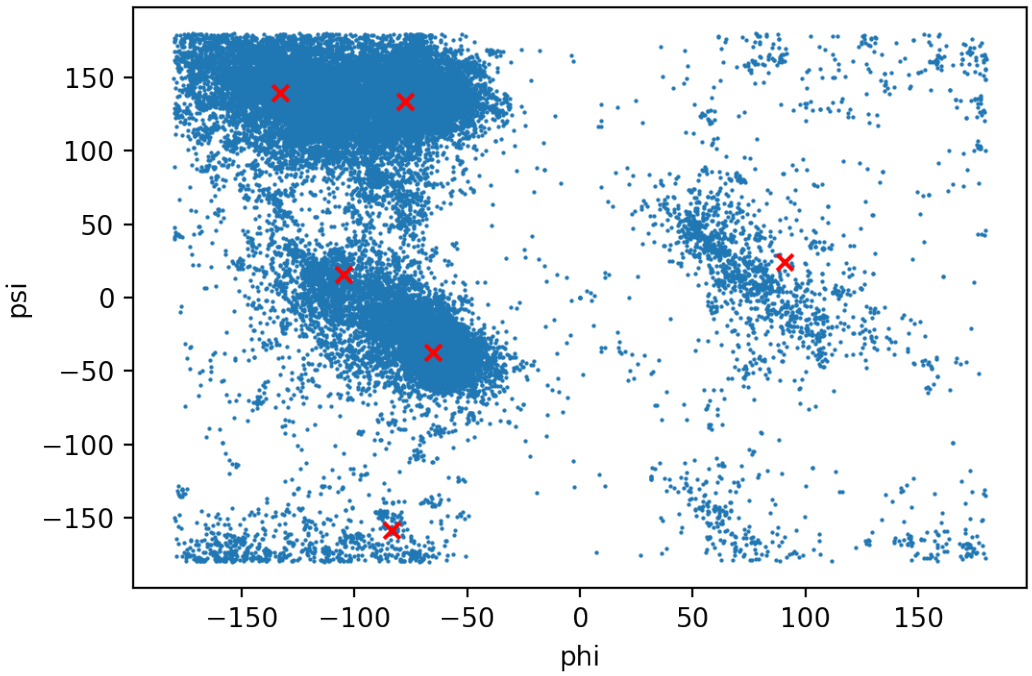

Figure 5: Cluster centers for K=6

To summarise, we believe a value of K=3 is the best choice for the data set as increasing it further does not create new groups but simply splits the already defined groups into new smaller groups.

2.1 Are the clusters reasonable?

According to Bioinformatics the clusters seem to be exactly within reason. The explanation for the clusterings lies in the inherent nature of the polypeptide chain that's most likely found in a special type of conformation, namely 'α-helices' seen in the middle-left, 'β-strands' observed in top & bottom-left and lastly 'turns' found in the middle-right. This type of formation, explained by the Ramachandran principle, is exactly what we observe in the clusters.

2.2 Shifting data for better results

When plotting different groups as different colours we can see that the large group at the top is split in two, as the data is periodic they should be part of the same group. As shown in the following figure, the highlighed data should be part of the red β group, not the blue α group.

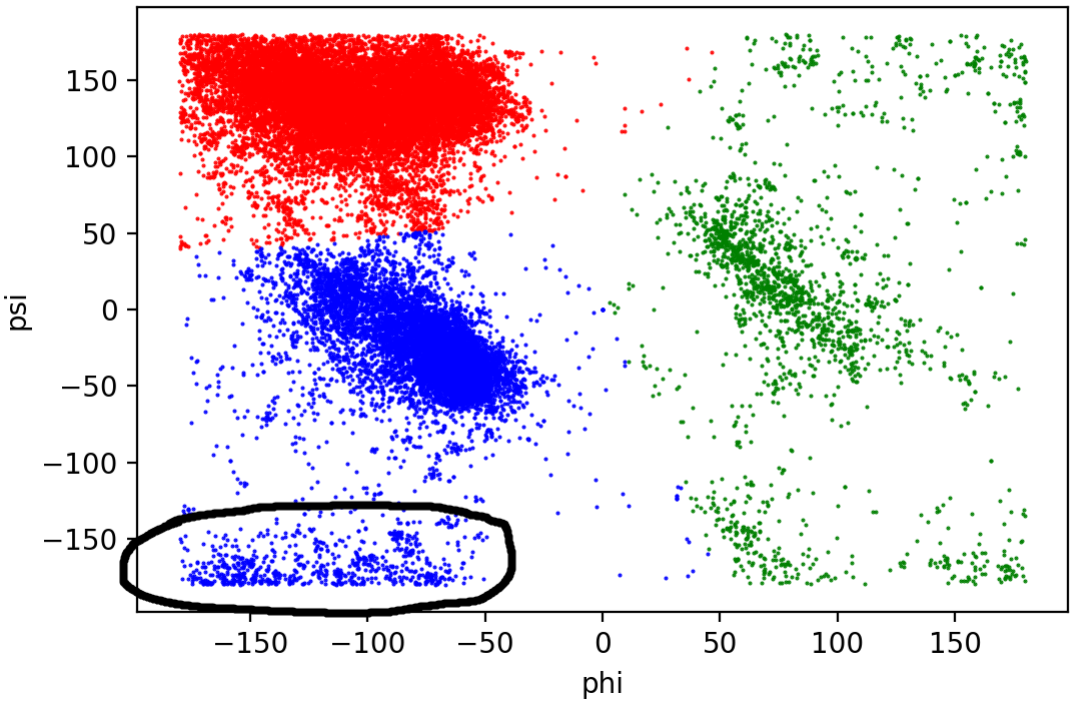

Figure 6: Split β group

This can be fixed by shifting the periodic psi data by -40 degrees. This result in the following correct grouping.

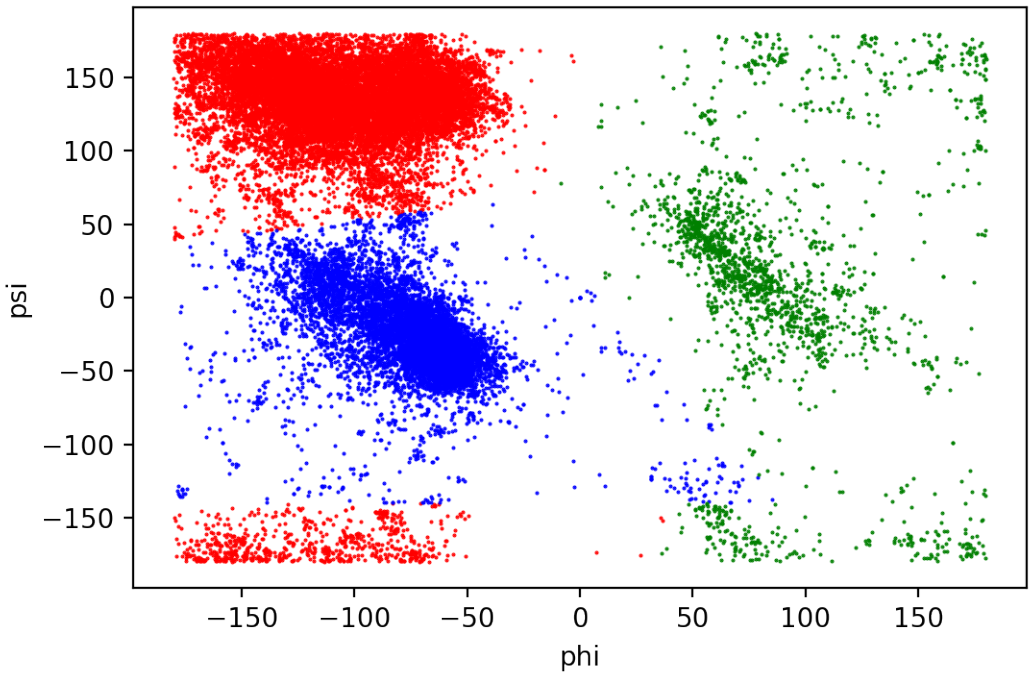

Figure 7: Complete β group

2.3 Validating the clusters

To validate the clusters we fist dropped a large amount of points and ran the K-Means algorithm with the same parameters on the new data. This resulted in the following 2 images where we can see that the clusters are almost identical, at least to the naked eye. (Except for the colors which should be disregarded)

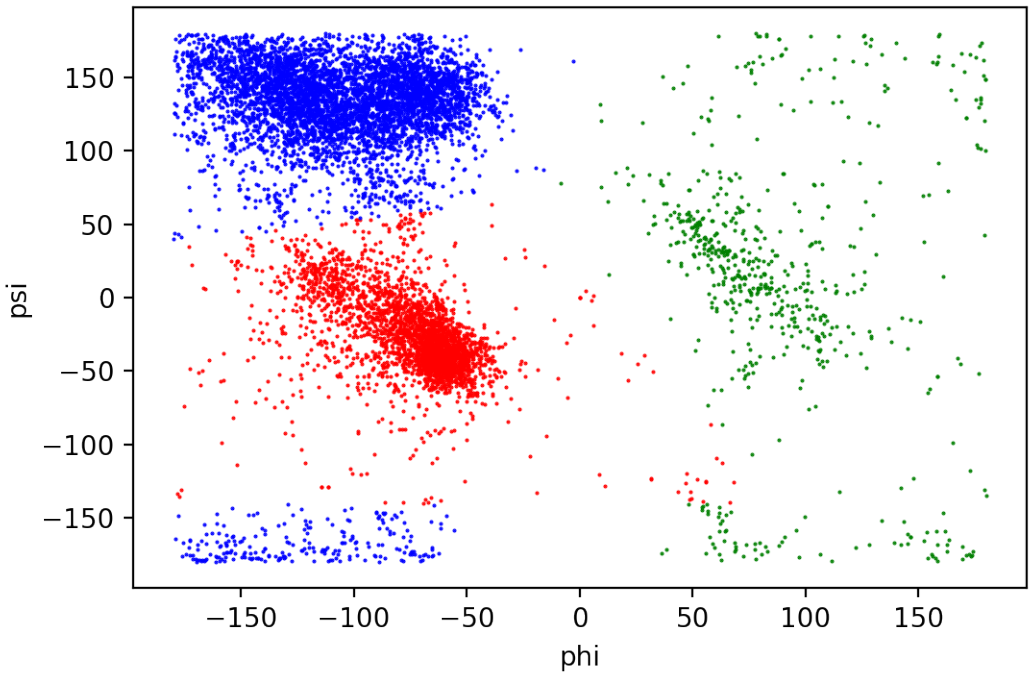

Figure 8: Clusters for K=3 with 20,000 randomly dropped points

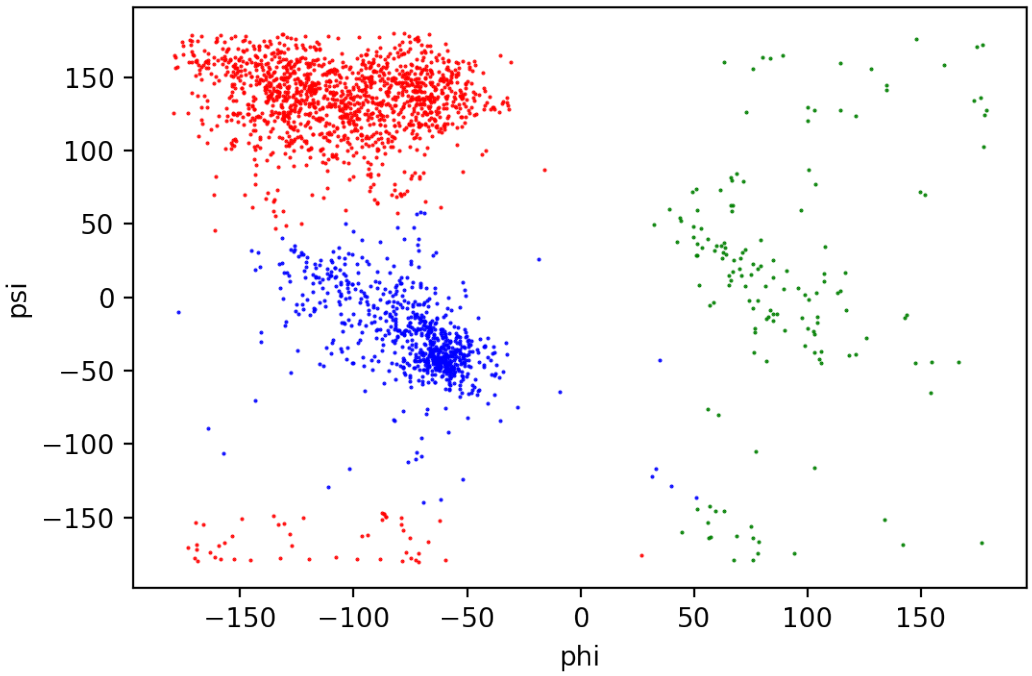

Figure 9: Clusters for K=3 with 27,000 randomly dropped points

For further validation, we also calculated the Silhouette Coefficient using the clusters resulting from the K-means method, which gave a score of roughly 0.68. This is a decent score, as the closer you get to 1 the better the clusters. A possible reason for the score not being higher might be because of the closeness of the *alpha* and *beta* clusters as well as outliers being counted for, bridging the gap between the clusters. As the clusters get closer, the algorithm might see it as the same cluster when, in fact, it's two separate ones (which it is in our case, discussed in the subsection "Are the clusters reasonable").

3 DBSCAN Clustering

3.1 Motivating minimum number of samples & epsilon

When choosing values for minimum number of samples and ϵ we needed an ϵ large enough to encompass sufficient data points but still ignore noise. As the clusters are densely packed with points the minimum sample count can be substantially large, this way only the dense parts of the data set would count as clusters. We used an ϵ value of 12 and a minimum number of samples of 160.

The resulting plot for the DBSCAN can be seen here.

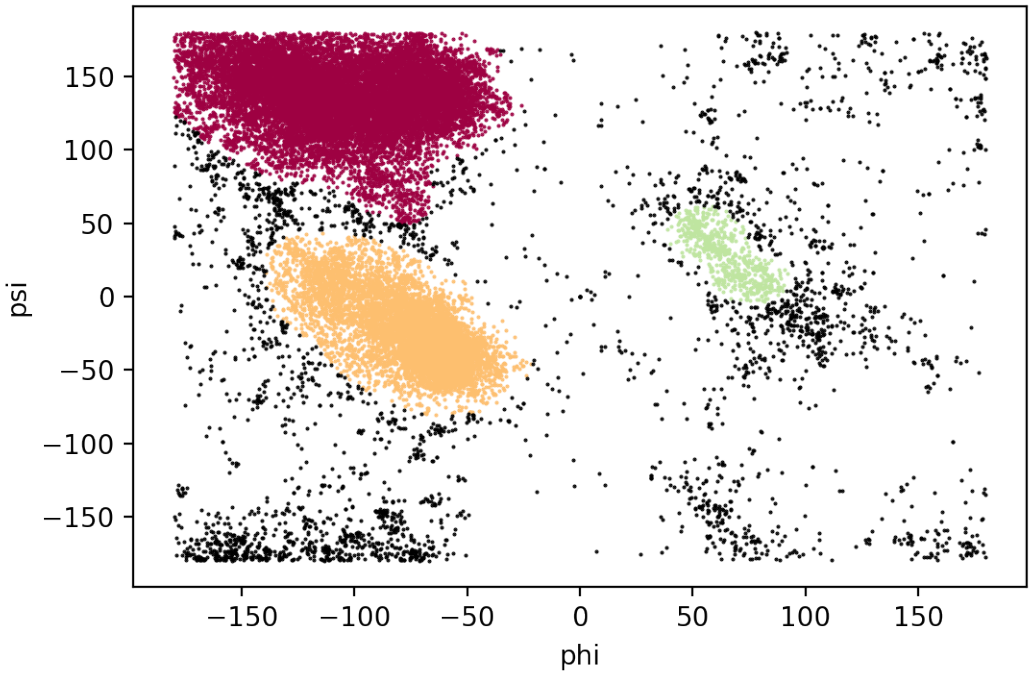

Figure 10: DBSCAN clustering of the phi and psi combinations

When using the same shifted data as in 7 we get a more accurate result. All further data in this section will be based around the DBSCAN of the shifted data (fig. 11).

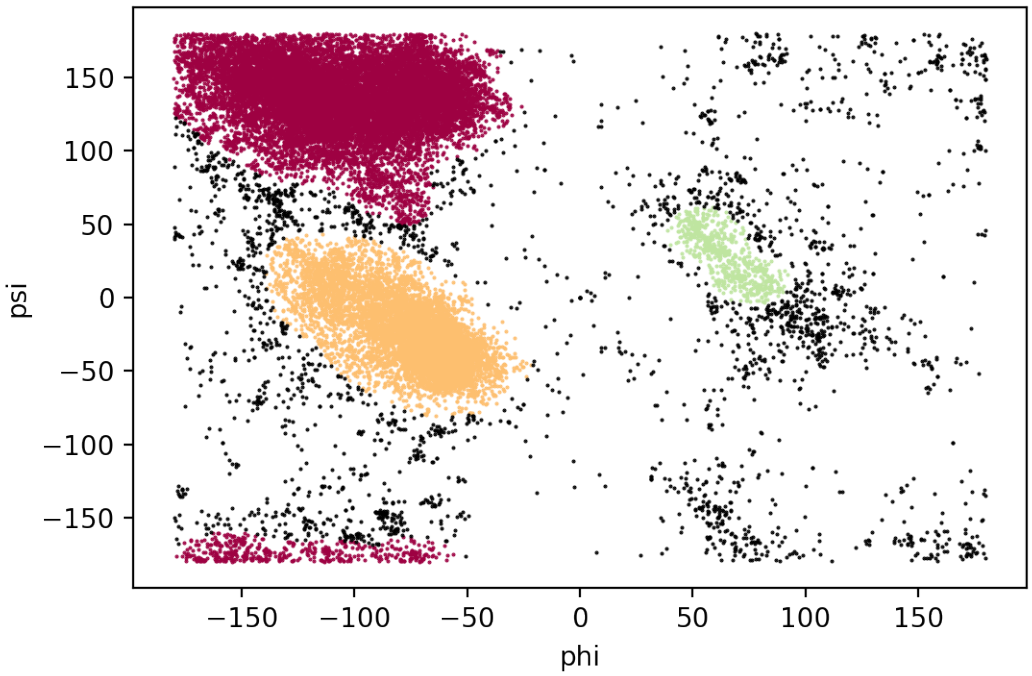

Figure 11: DBSCAN clustering of the shifted data

Above you can observe the scatter plots highlighting the clusters found as well as the outliers, represented by black points, using the DBSCAN method. This method yielded 3 clusters, the aforementioned α, β and *turns*, and 2672 noise points, see figure 11.

3.2 Residue type of outliers

Plotting the residue type of the noise points we can see that Glycine is by far the most common one (fig. 12). According to Biozentrum, University of Basel "Glycine has no side chain and therefore can adopt ϕ and ψ angles in all four quadrants of the Ramachandran plot. Hence it frequently occurs in turn regions of proteins where any other residue would be sterically hindered". From this we can deduce that Glycine being the most common outlier is a reasonable result.

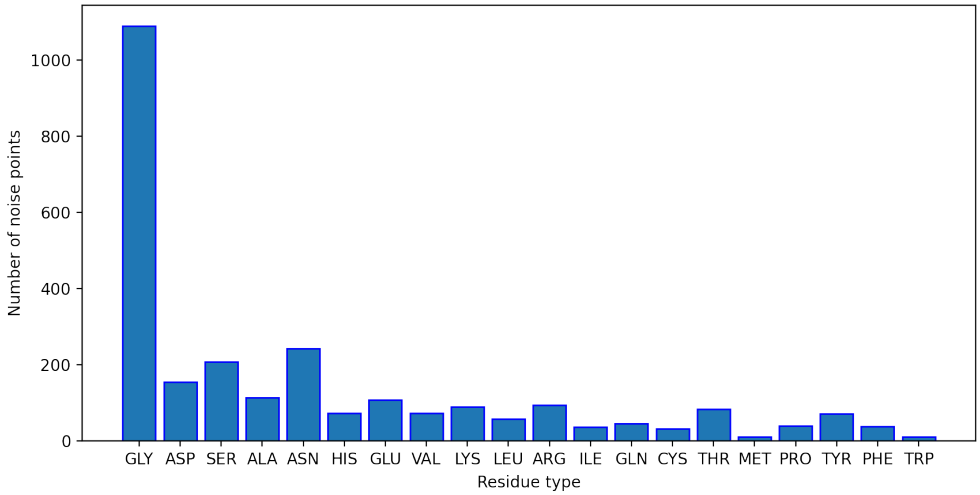

Figure 12: Residue type of noise points

3.3 Comparing the result of K-means against DBSCAN

When comparing the two methods for clustering we can see similar groupings, however, using DBSCAN the clustering takes outliers into account, this is not the case for K-means. This is a major flaw in the K-means method as it skews the data by potentially dragging the centroids towards outliers that should not be part of the cluster, see figure 3. Conversly, DBSCAN uses a density-based clustering algorithm that treats low-density areas as noise and for our use-case this is the more accurate version.

3.4 DBSCAN Robustness

DBSCAN is unstable when it comes to small changes changes in ϵ and minimum samples. The reason for this is the same as described in the subsection "Motivating minimum number of samples & epsilon". When changing the ϵ we might encompass too much and drag in outliers or encompass too little to create the right clusters and the same reasoning can be applied to changing the minimum amount of samples in the neighbourhood. This is clearly shown by the following figures where we have altered the parameters only slightly, see the figures captions for the change in values. Worth mentioning is that how much you can change the parameters ϵ & *minimum samples* differs from dataset to dataset.

Note: Starting values for ϵ and min samples is 12 and 160

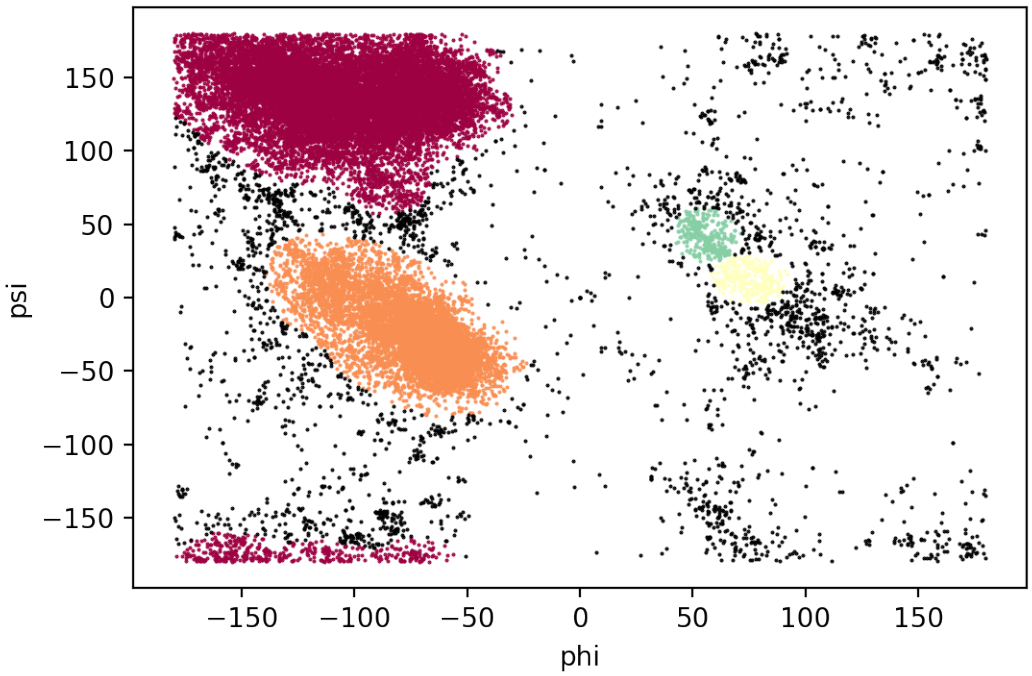

Figure 13: DBSCAN with minimum samples increased by 10

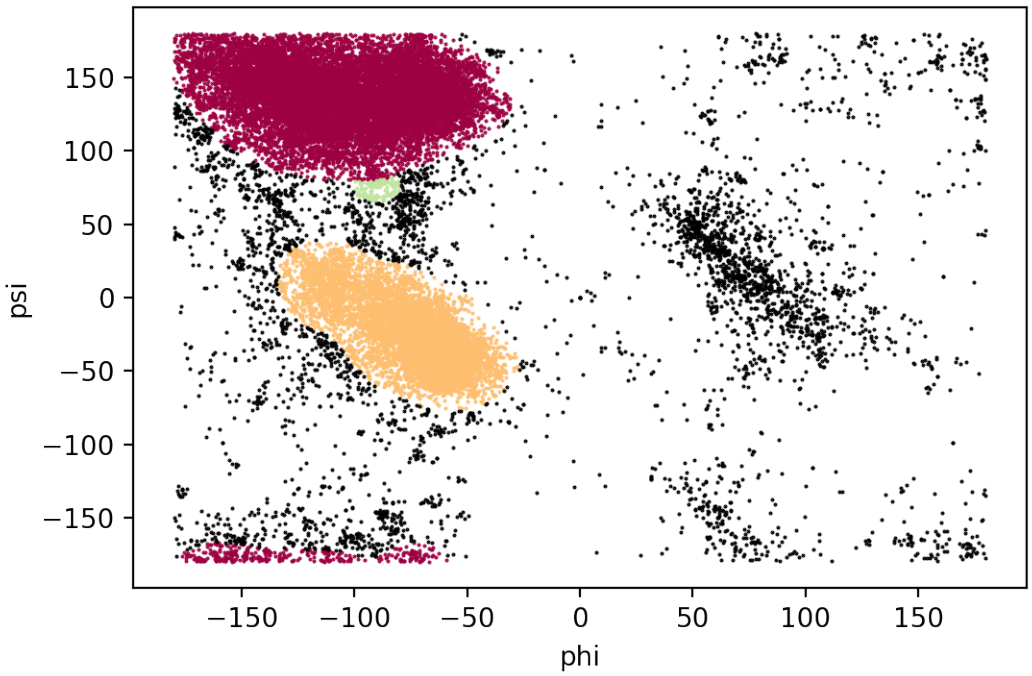

Figure 14: DBSCAN with ϵ decreased by 2

4 Stratified data clusters

Stratifying the data based on amino acid residue type and looking specifically at PRO and GLY types we get vastly different results.

When looking at PRO we can see that there are very few outliers, which is also shown in the barchart for outliers (fig. 12). Most of the amino acids of this type are found in the large β and α groups and none are found in the group to the right.

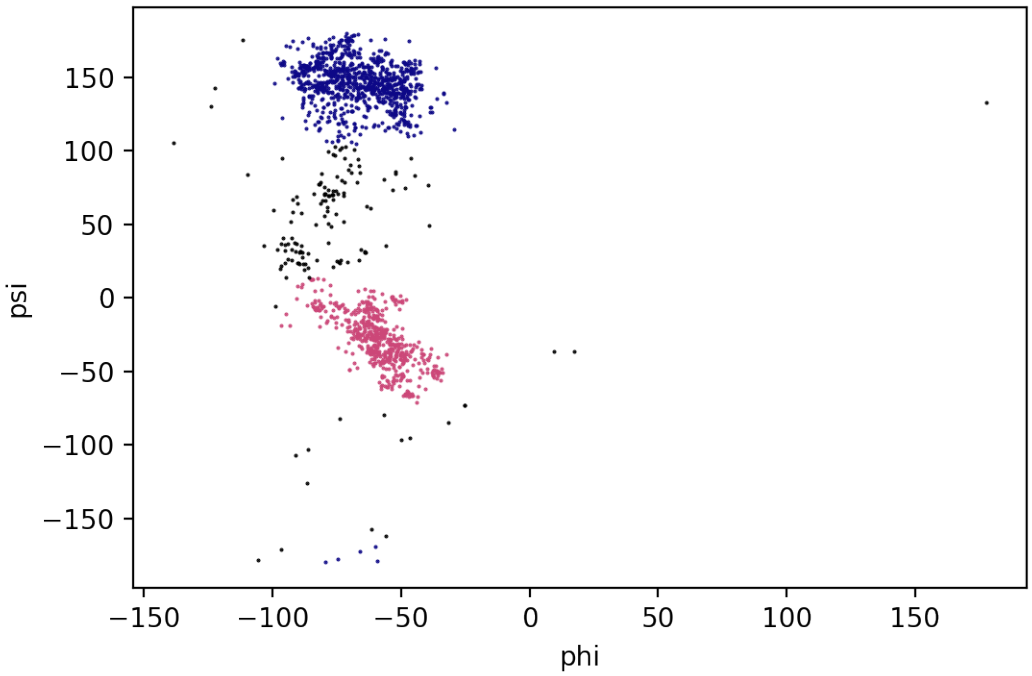

Figure 15: DBSCAN of PRO amino acid residue type

When looking at GLY on the other hand, the complete opposite is true. GLY can be found all over the Ramachandran plot with the highest concentration in the group to the right. This clearly shows why GLY is the most common outlier, why the behaviour of GLY is so different from the rest of the amino acids is explained in the section on "residue type of outliers".

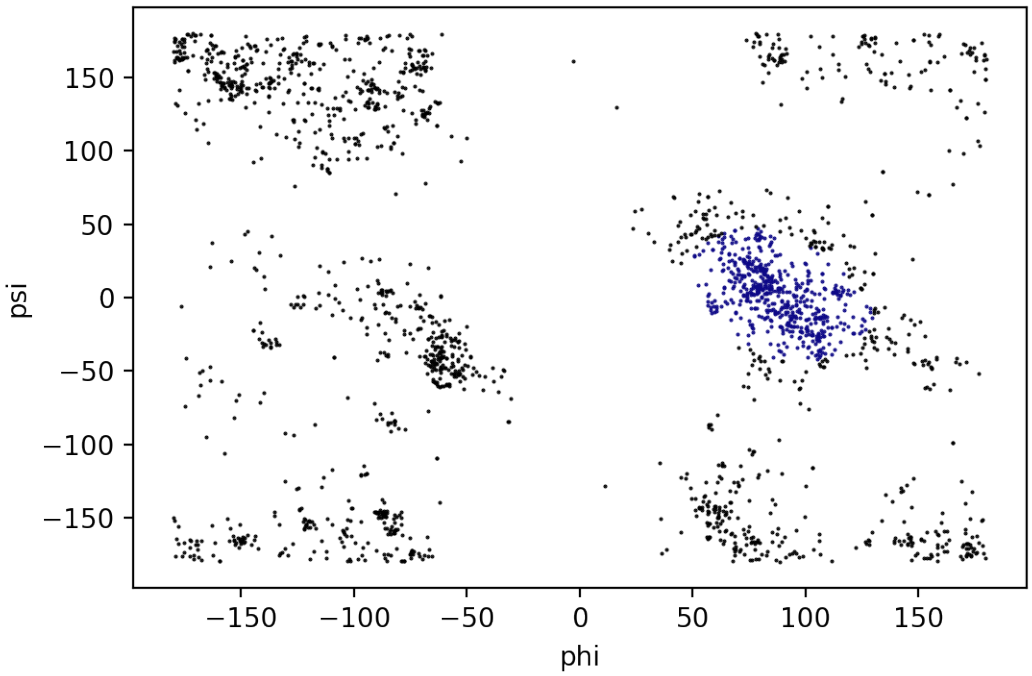

Figure 16: DBSCAN of GLY amino acid residue type