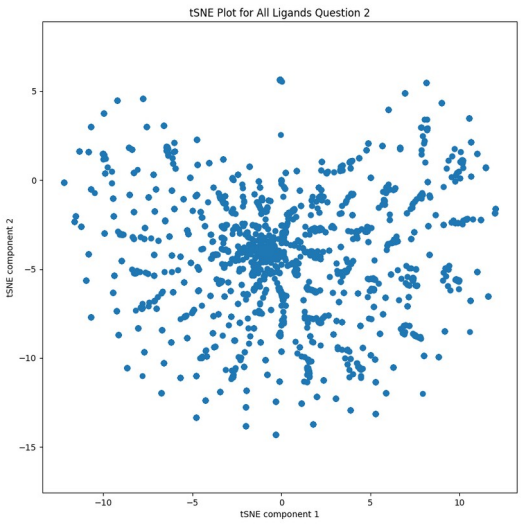
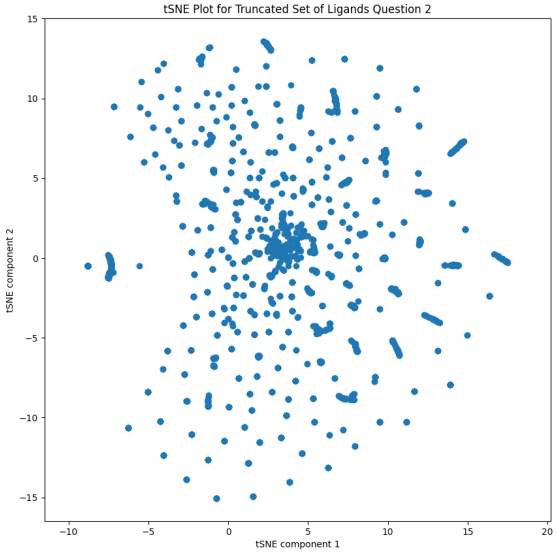
Benjamin Wheeler
BMI 203 Project 2 Write Up

Part 2 Questions

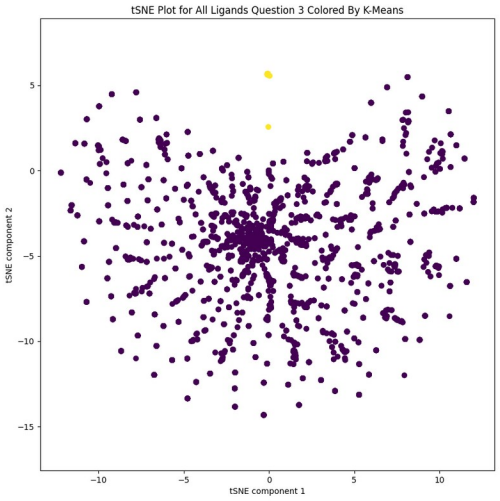
1.) My two different clustering algorithms employ slightly different similarity metrics, but at the assignment level of which cluster a ligand belongs to both use essentially euclidian distance. In the hierarchical algorithm, euclidian distance is used pretty much as normal. This was used given the straightforward implimentation and interpretation when considering how “close” two ligands are in high dimensional space.
The partition clustering algorithm as mentioned makes assignments based on euclidean distance as well. However, in this case these distances are calculated based upon the similarity matrix. This is the idea that the similarity matrix serves as an artificial (ligand_num x ligand_num) space defined by continuous values. The similarity matrix is calculated using the tanamoto coefficient. This allows centroids to be easily calculated and implemented in this continuous space.

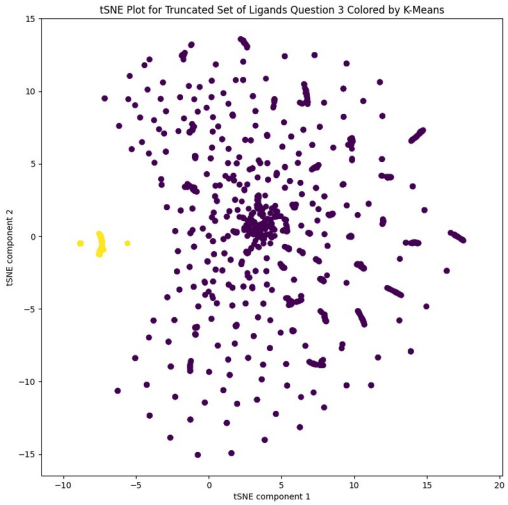
2.) tSNE plots for all the ligands provided and the truncated set used for hierarchical clustering



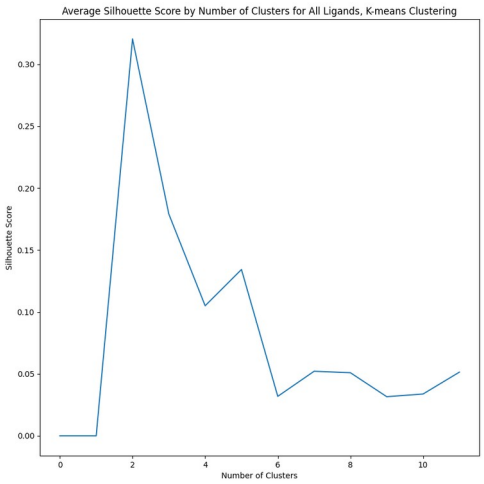


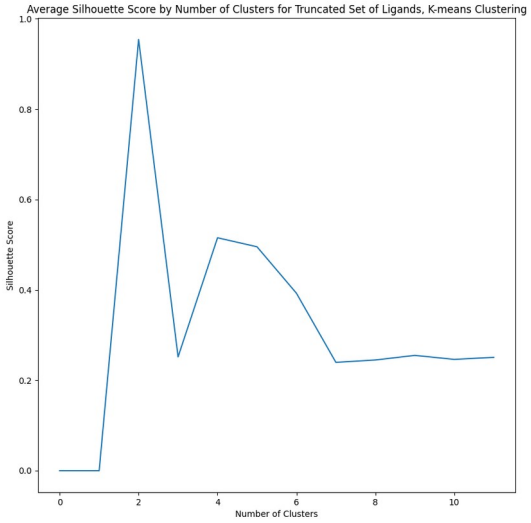
3. Same tSNE plots as above however the dots are now color coded by the two clusters present



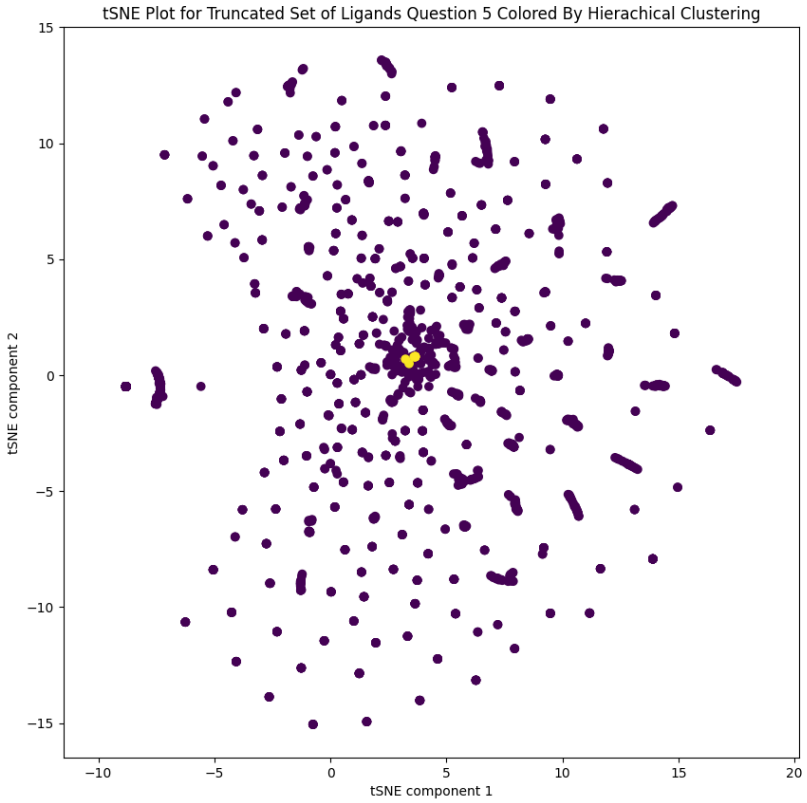


4.) My partition clustering algorithm is a vanilla implementation of K-means. The choice of sticking to vanilla K-means was a practical decision. I first wrote the vanilla k-means with an intention to upgrade the centroid initialization to something more like a K-means++. However, when I applied K-means in test runs to the data, the strange shape of the data became apparent. By this I mean that about 1/3 of the data set are highly similar and tightly clustered (by visual inspection in t-sne). The remaining 2/3 of the data set are diffuse; they are both far away from the tight cluster and each other. So in implementing vanilla k-means the solution often quickly converges to 2 main clusters. The main issue is that some ligands have identical finger prints which means after updating cluster assignments some clusters will be empty. K-means++ implementation would reduce the frequency of this occurring, but would be more complex to implement rather than the solution I took. To address this I re-initialize centroids of empty clusters to random data points until the problem is solved. I mentioned above that visual intuition with the t-sne plot indicates the presence of two clusters. This is further supported by calculating the average silhouette coefficient across all data points for a given number of clusters. This is plotted below and shows that this peaks nicely at 2 clusters. By this metric k=2 is ideal for the vanilla implementation of k-means I have done here. This algorithm is sensitive to initial conditions, but given the data this seems to be most relevant when k >2. In this case the initial centroid choices impact the much smaller auxiliary clusters. In this particular case the two large main clusters are very stable.

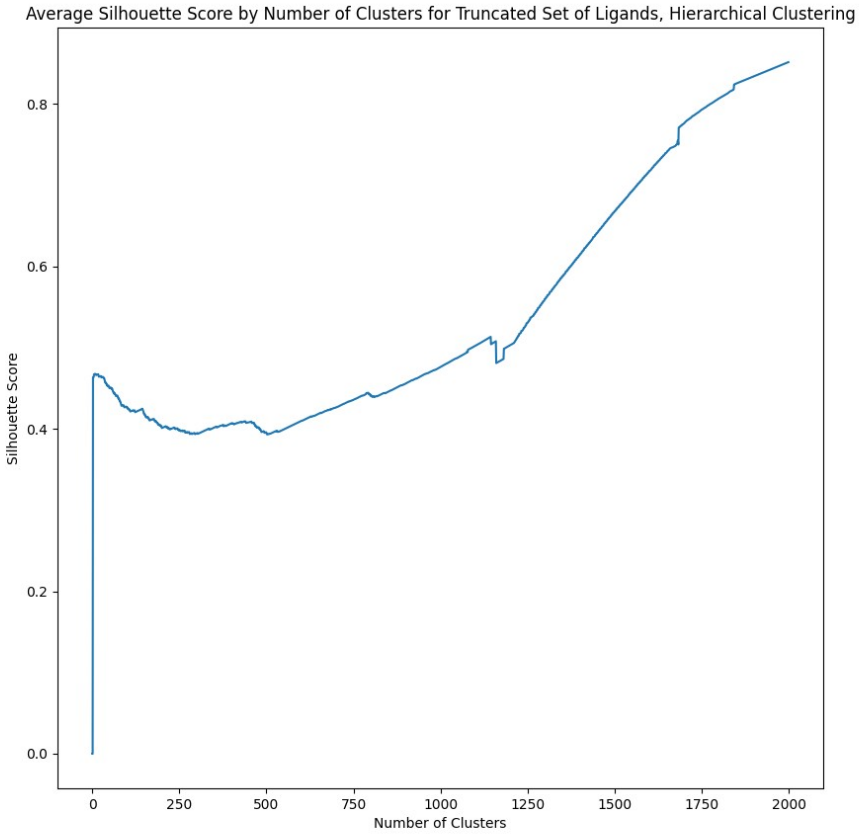




5.) Same tSNE as above for the truncated ligand set, now the ligands are color coded by the hierarchical clustering cluster assignment.



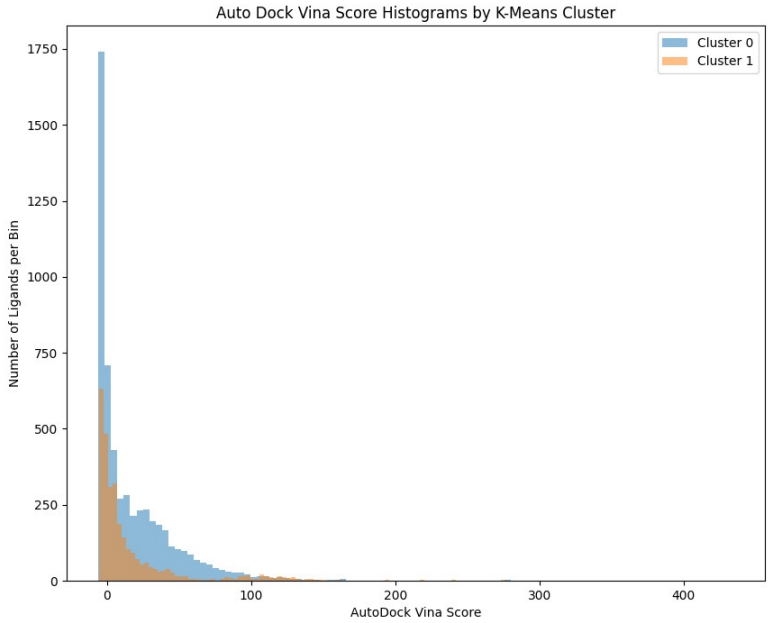
6.) The hierarchical algorithm I implemented is an agglomerative, bottom up approach using a single linkage joining criteria. I chose to use single-linkage criteria for one in that it is slightly faster than other linkage criteria such as average linkage. This is still incredibly slow because it scales O(N^3). Given the time to run the algorithm, I have not compared it to complete linkage. Given the sparse, high dimensional nature of the data, I made the assumption that single-linkage will perform better. This could make many pairs of ligands so far apart that complete linkage would perform poorly. My hypothesis is that single linkage would do a better job handling this scenario. This hypothesis has not been tested and is outside the scope of this write up. This clustering method is deterministic in that it is not sensitive to initial conditions and there are no stochastic elements to the algorithm. I determined the number of clusters to use in a similar manner as I did for the k-means algorithm. I tested many different numbers of clusters to use and calculated the average silhouette score for the ligands clustered. The plot below illustrates the complex nature of determining the ideal number of clusters to use for hierarchical clustering of this data set. If all possible cluster numbers are considered then we can observe a roughly 0.46 silhouette coefficient for the small cluster numbers. This roughly decreases with number of clusters until 250 at which point the silhouette coefficient climbs again, with a wide peak at approximately 2000 clusters. Given that 2800 ligands were clustered, this number of clusters provides very little added value from evaluating the ligands individually, the local maximum at the beginning was chosen instead. This corresponds to 8 clusters. This behavior likely underlies that the molecules are highly diverse and do not possess a natural hierarchical relationship to each other.



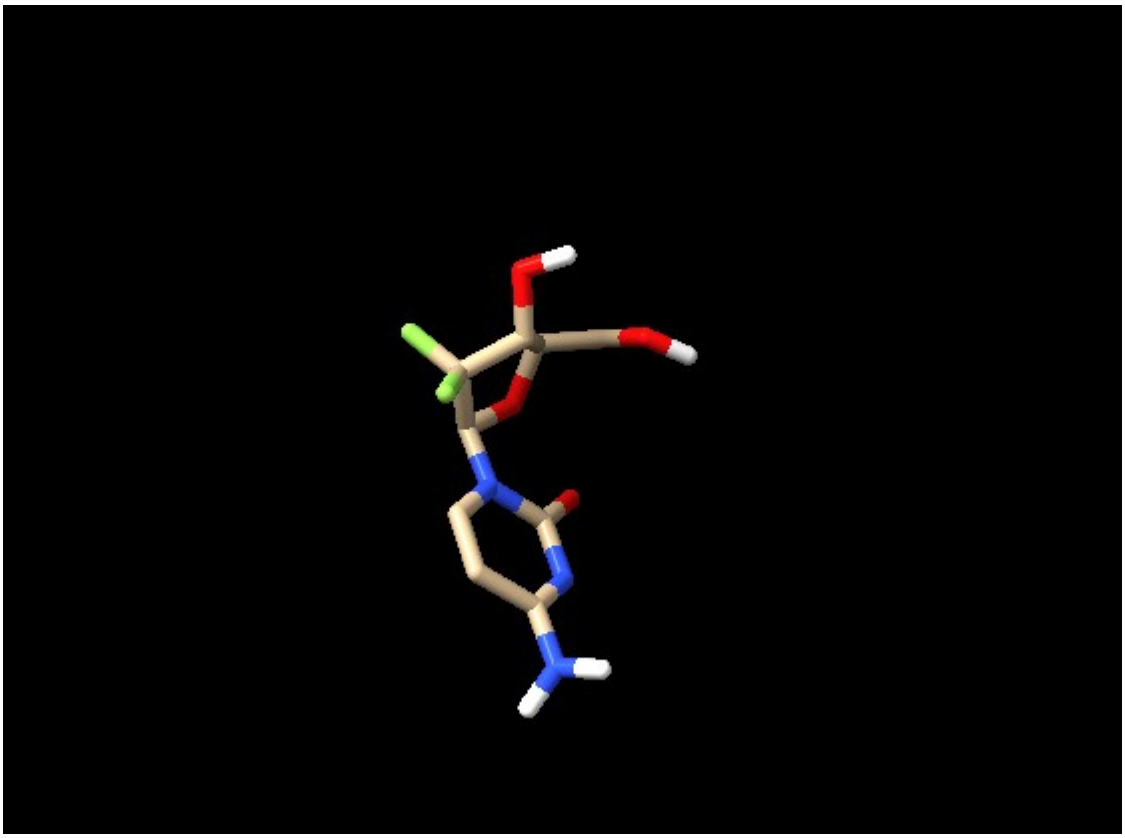
7.) To evaluate the two clustering algorithms I again employed the silhouette score from the ideal number of clusters chosen for each algorithm. I chose the silhouette score because it synthesizes both the cohesion of a single cluster and the separation of the clusters. This is useful because the hope is that clustering can reveal how chemical similarity relates to binding affinity for a given protein (complex). If the clusters were not separated well, we would have difficulty finding this relationship because the clusters were not substantially different. If the clusters were not cohesive, then the intra-cluster heterogeneity would introduce too much noise to make conclusions. By this metric the k-means clustering with just 2 clusters performs the best with a silhouette score above 0.9, substantially above 0.46 for the hierarchical clustering algorithm.

8.) I compared these two clusterings on the down sampled set 2841 ligands. By using the jaccard coefficient, I obtained a value of 0.5422. I chose the jaccard coeffecient because, given the nature of the problem, mutually classifying two ligands as different did not seem of as much value. Given the value of 0.5422, it would suggest that the clusterings are roughly 50% identical. This makes intuitive sense given that the k-means clustering only has two predominant clusters. However, the ligands are not partitioned evenly into those two sets, so the hierarchical clustering probably does a slightly better than 1 single giant cluster.

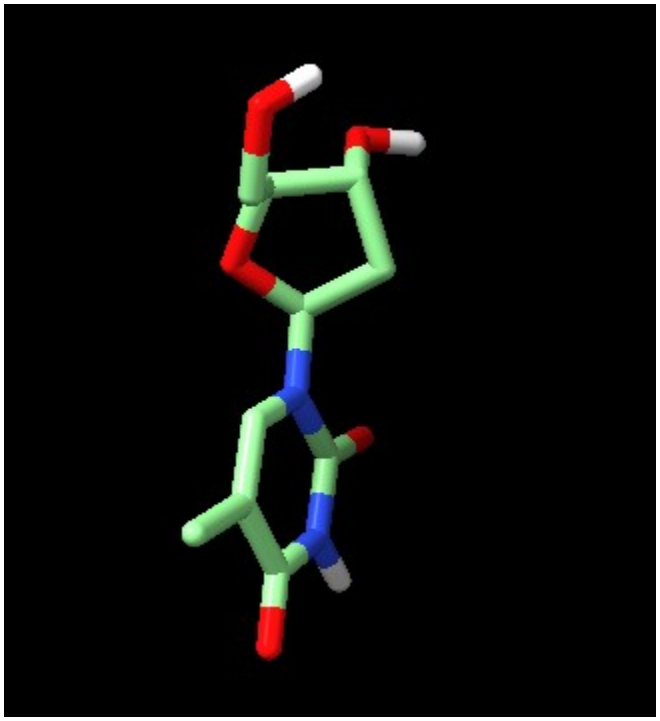
9.) The scores across the two clusters in my K-means implementation are quite widely distributed. Cluster 0 appears to be more widely distributed than Cluster 1. In particular, cluster 1 really falls off after a score of 50 while there is a grouping of cluster 0 ligands around this mark. It should be noted, however, that the mode of each cluster is less than 0. Additionally, cluster 1 has a lower average score compared to the average in cluster 0. It would actually be quite surprising if the members of a cluster had very similar scores. The clusters indicate some shared presence of a chemical moiety. The local structure of the molecule around this chemical moiety and the global geometry of the molecule are not considered. These will play a large role in getting the right chemical moieties to interact with the proper residues on the protein, so the inability to represent that in our chemical signatures will make it difficult to identify groupings that have similar scores.



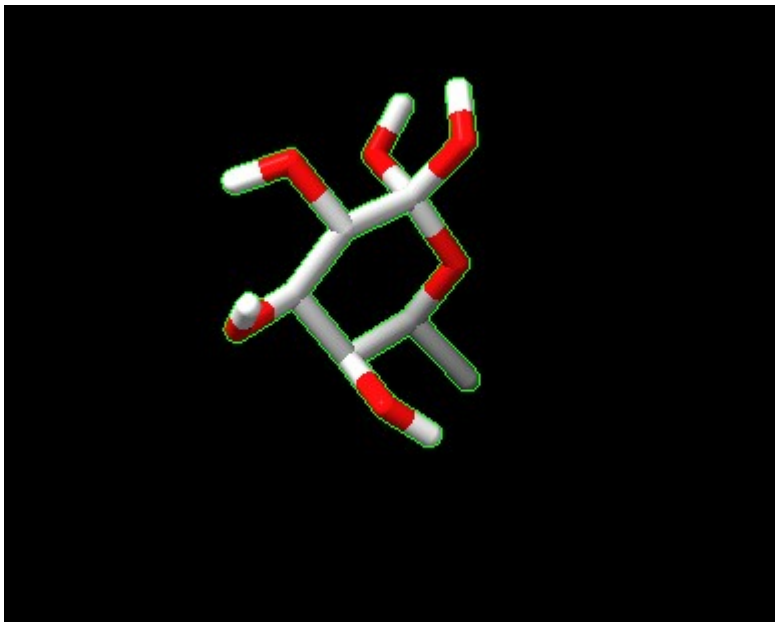
10.) Top cluster 0 Ligands:
2019



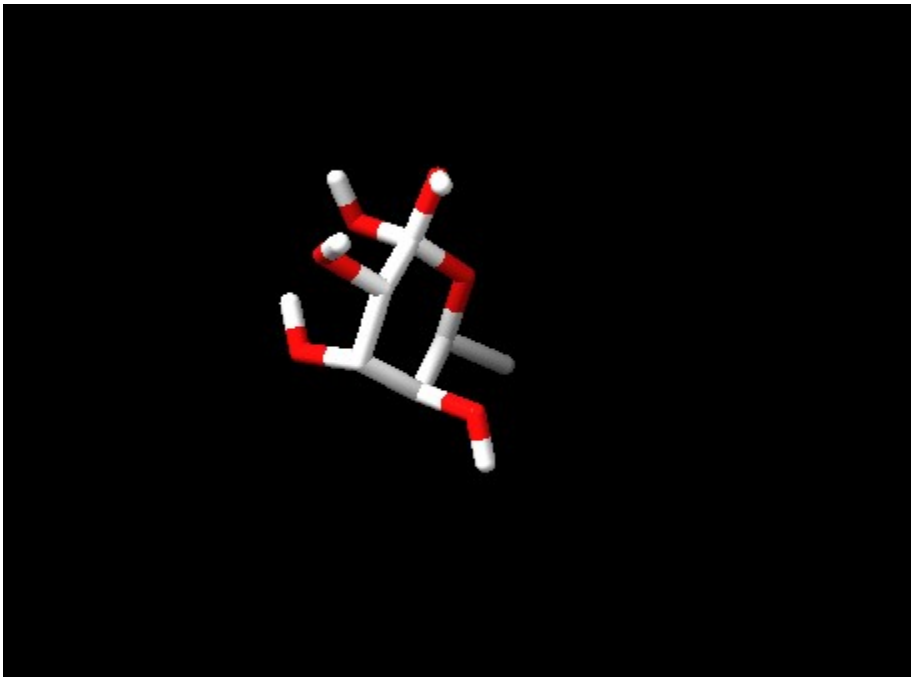
1628



740

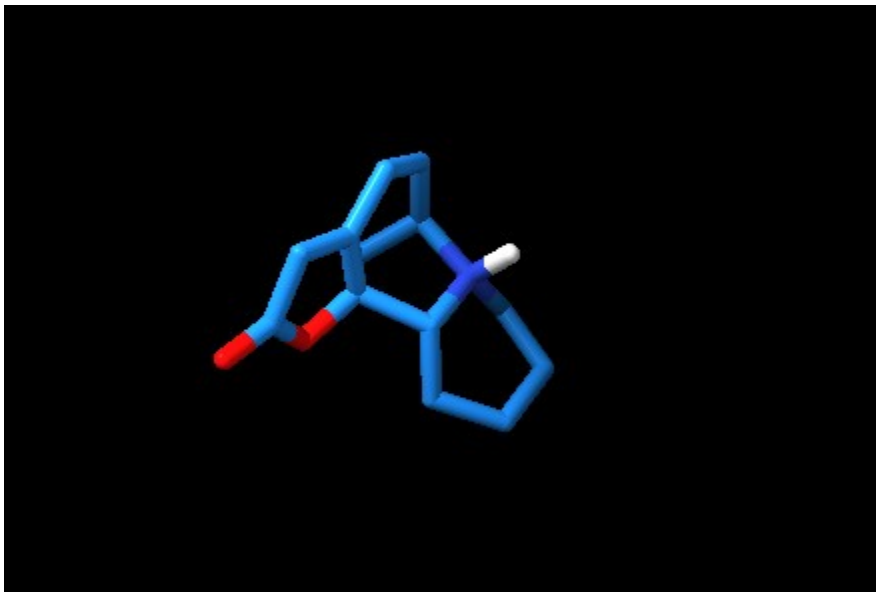


736

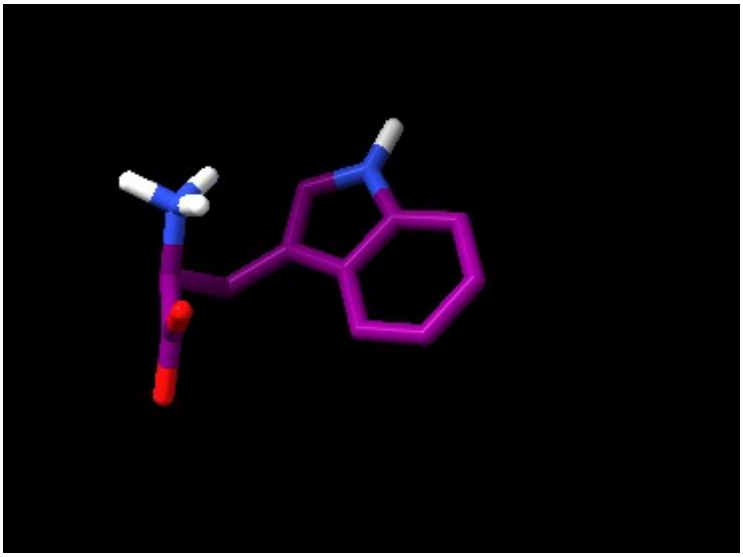


Top cluster 1 Ligands

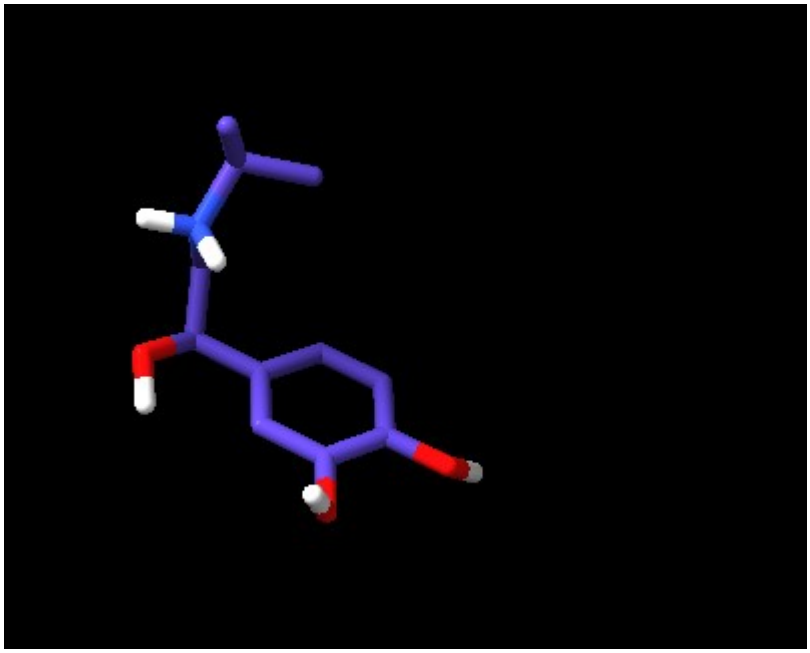
1267



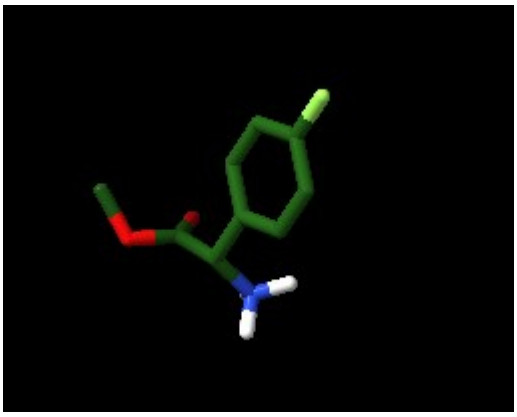
1038



1199



810



I would preface this section with the fact that I am not a medicinal chemist by a long shot. However, I broadly feel these molecules to be quite diverse. There are some elements that seem shared between pairs of the molecules here, but not across all the ones selected. For instance, in cluster the alcohol groups at the head of the molecule seem to have some shared properties. Molecules 740 and 738 are actually highly similar, and might just be the same molecule with different stereochemistry. These two molecules are simply a ring decorated with many alcohol groups. This might suggest the importance of this group in cluster 0. Cluster 1 molecules seem to be more diverse, with many different backbone and other moieties present.