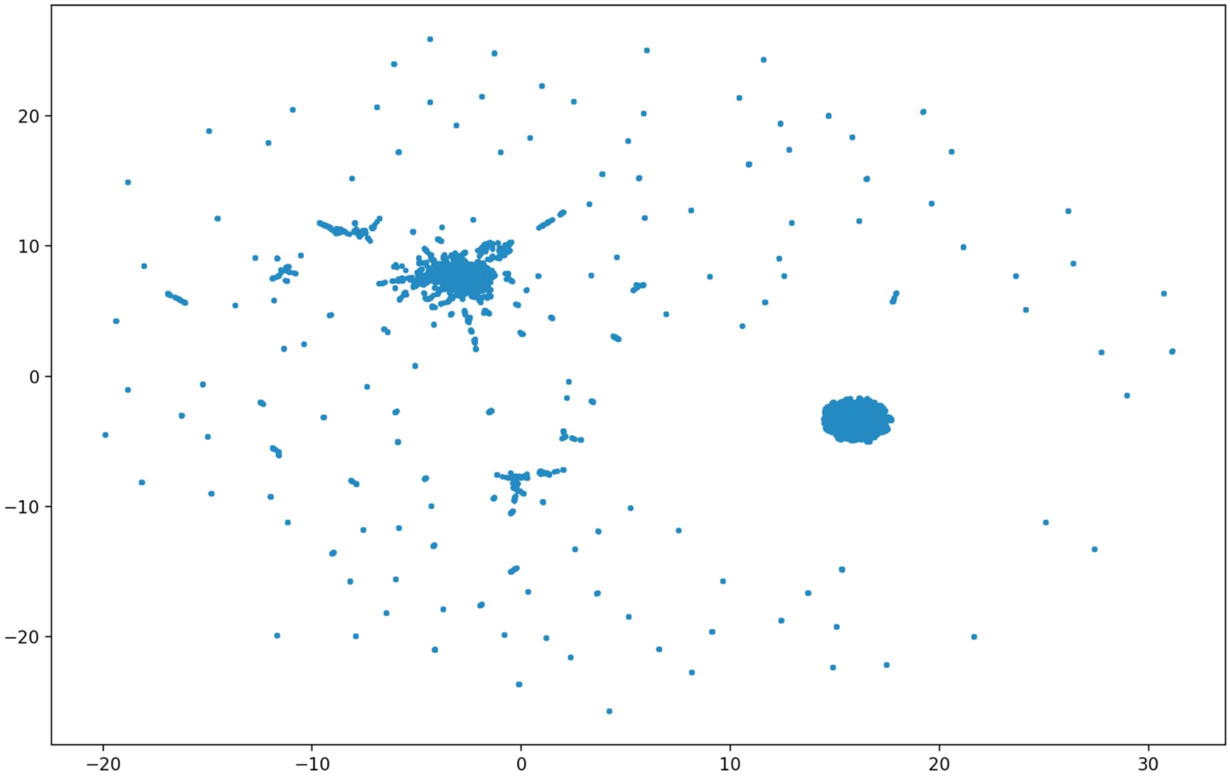
Ryder Easterlin
February 19, 2021
Project 2 Report

1. **Explain the distance metric you utilized to calculate the similarity/dissimilarity between small molecules.**

I utilized the Tanimoto similarity between bit-vectorized molecular fingerprints to assess similarity between two small molecules. To make Tanimoto similarity into an appropriate distance function, I simply took (1 – Tanimoto Coefficient) as the distance between the two bit-vectorized fingerprints.

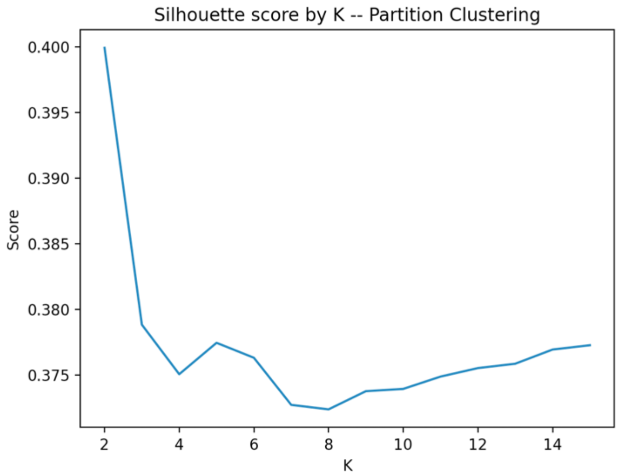
2. **2. Use a dimensionality reduction algorithm (PCA, t-SNE, UMAP, etc) to generate a 2D visualization of the small molecule dataset. Each point should represent a single molecule.**

I used UMAP to reduce the feature space, and each ligand is plotted below.

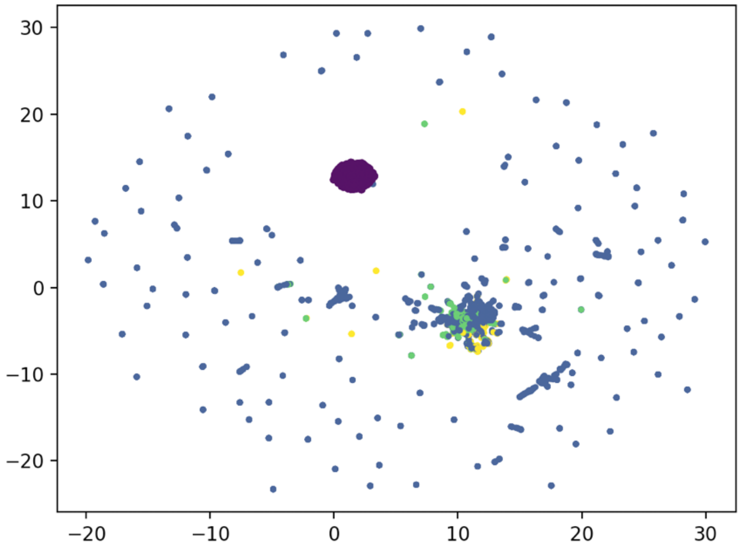


3. **Cluster the small molecules using your implementation of a partitioning clustering algorithm. Visualize this clustering by coloring clusters on the 2D visualization generated in question 2.**

I implemented K-means clustering as my partition clustering algorithm, so I need to determine the best value of K before visualization. Based off of the above visualization, I would say there are approximately 5-10 main clusters in the data. So, I will try a range of K between 2-15 and assess the quality of each clustering by its mean silhouette score, visualized below.

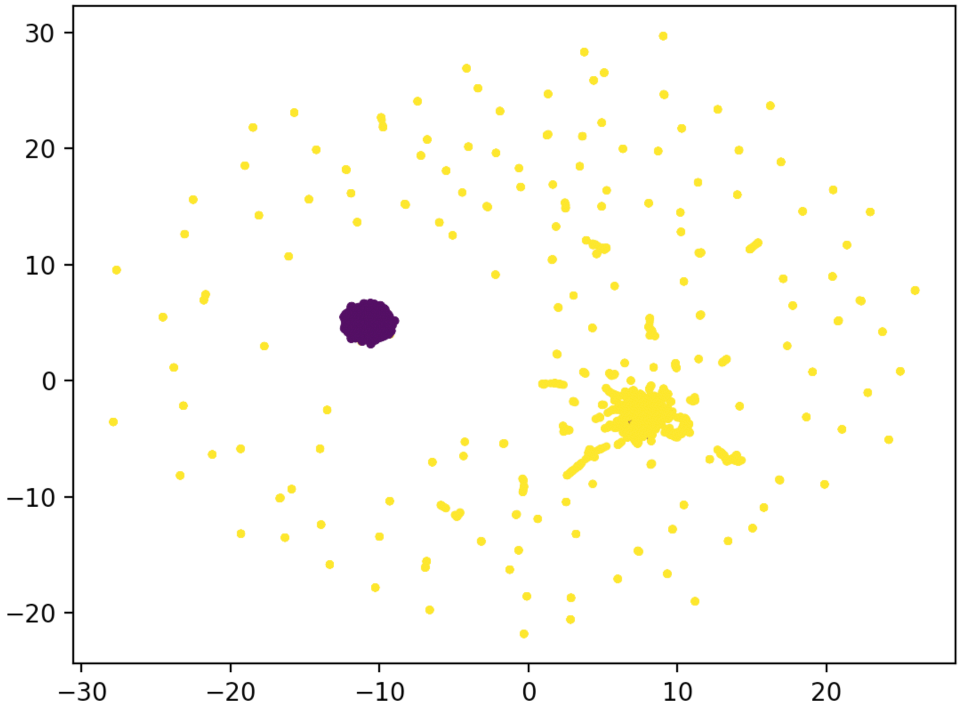


The best scoring performing K-values are 2 and 3, but I don't really think a K-value of two matches the data that well (at least according to the UMAP visualization, so I will instead use a K of 5, since 5 is the best K-value besides 2 or 3.



Unfortunately, the Kmeans clustering doesn't match up all that well with the clusters that we can see in the UMAP plot. That said, this may have more to do with the low-dimensional representation generated from UMAP not being a perfect representation of the higher-dimensional data that was clustered. This may result in some degree of non-overlap between the real results and what we expected based off of looking at the UMAP plot naively.

Since the above plot was a little disappointing, I retried with K=2 and the resulting plot is below.



4. **Explain your choice of partitioning clustering algorithm. Is it sensitive to initialization conditions? How do you select the number of clusters?**

As touched on above, I use the K-means clustering algorithm. K-means is known to be sensitive to initialization conditions, specifically the initial placement of cluster centroids in the feature space. I use a simplified version K-means++ in my implementation to attenuate this issue. Instead of sampling from a probability distribution of distance from each centroid, each new centroid is simply chosen as the point that is the furthest from its closest existing neighboring centroid.

Selecting the best value of K is another challenge in K-means clustering. In order to decide on K, the qualities of clusterings using different values of K should be compared with some kind of quality metric. I use the mean silhouette score of all data points in a clustering as the quality metric, and choose the value of K that maximizes this score. Inherent structure in the data should be considered also, and this can be queried be transforming the data into a low-dimensional feature space and visualizing it in this space.

5. **Cluster the small molecules using your implementation of a hierarchical clustering algorithm. Visualize this clustering in the same way as question 3.**

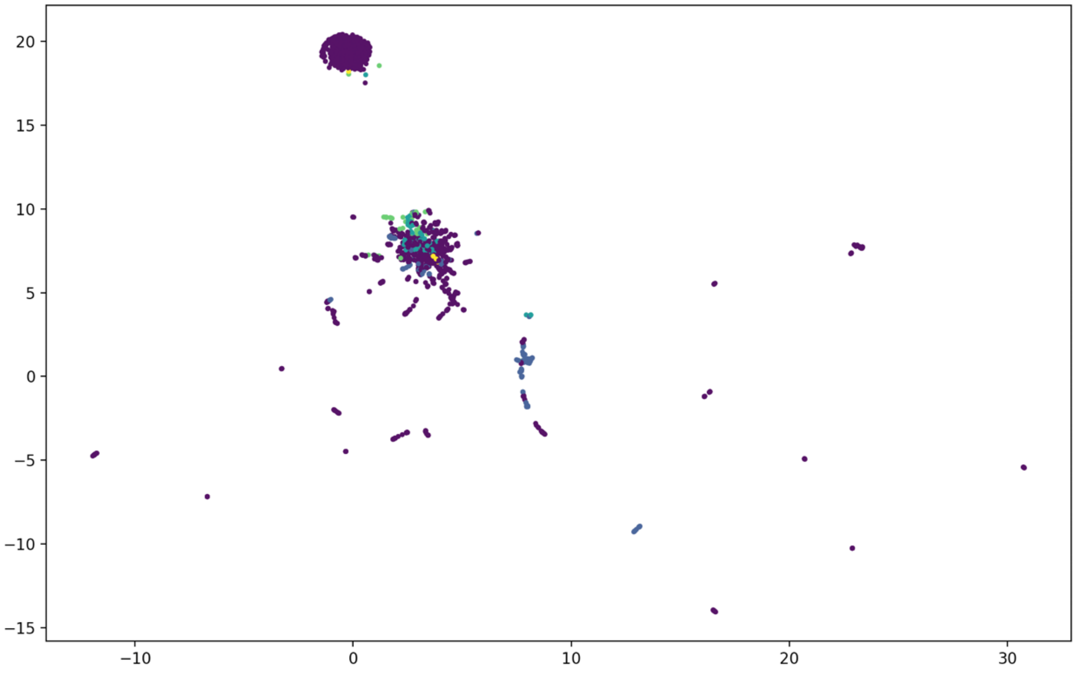
Because hierarchical clustering has cubic worst-case time complexity, I will be performing hierarchical clustering on a subset of 2000 randomly selected ligands from the main dataset, and for future questions will be comparing the performance of the hierarchical and partition clusterings on the same subsets for consistency.

I generated a clustering with 5 clusters, since that is the value of K I used above.

6. **Explain your choice of hierarchical clustering algorithm. Is it sensitive to initialization conditions? How do you select the number of clusters?**

I implement agglomerative hierarchical clustering, which is not particularly sensitive to initialization conditions since each observation starts in its own cluster, and each iteration joins the two closest clusters with a non-probabilistic metric. Cluster "closeness" is defined by a linkage criterion, and results are dependent on which linkage type is used. I use complete linkage in my implementation, which defines the linkage between two clusters to be the distance between the two furthest points (one in each cluster). I tried single and average linkage as well, but complete linkage had the most consistent results.

To determine the number of clusters, one would ideally perform a similar protocol as I described for partition clustering. However, since hierarchical clustering is so slow this was not feasible for the scope of this project. Therefore, I used 5 clusters, since this is the number I chose for K-means clustering and using the same number makes it easier to compare the two clustering types in the following questions.



7. **Evaluate the quality of both clusterings using your implementation of a clustering quality metric. Explain your choice of quality metric. Which clustering performed 'best' according to your metric?**

I used the mean silhouette score of all data points in a clustering as that clustering's quality metric. Silhouette scores have a range of [-1,1], with 1 indicating a near perfect clustering, 0 indicating more or less random assignment of ligands to clusters, and -1 indicating closeby points being assigned to separate clusters.

K-means silhouette score: 0.3685894372318187
Agglomerative silhouette score: 0.1280296221814476

8. **Compare the two clusterings using your implementation of clustering similarity. How similar are the two clusterings using this function?**

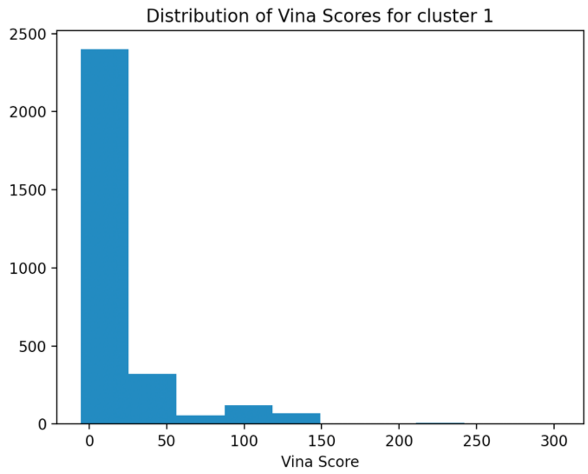
I use Tanimoto similarity as my clustering similarity function, inspired by the function described here: https://rdrr.io/cran/clusteval/man/jaccard_indep.html. This metric has range [0,1], with 0 indicating complete dissimilarity, and 1 being a perfect match.

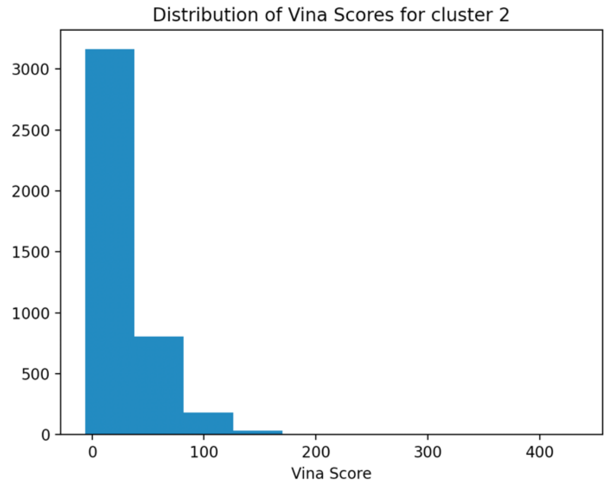
Similarity: 0.2860788639229661

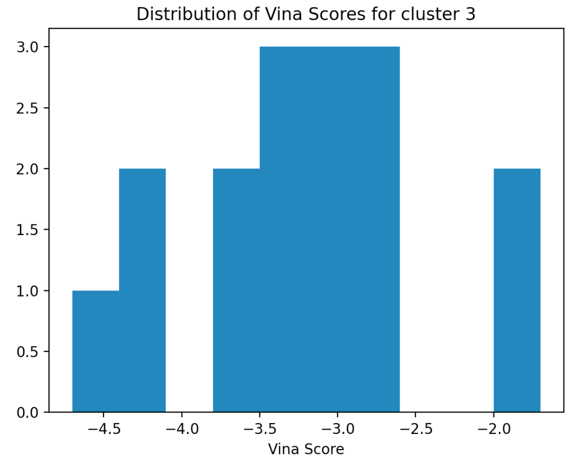
These clusterings are fairly distinct, with the partition clustering performing the best.

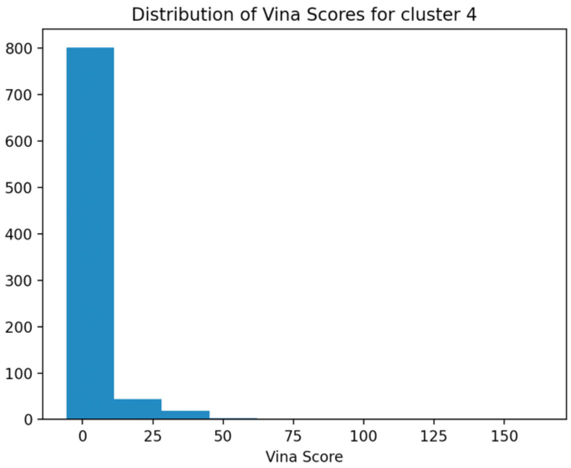
9. **For the “best” clustering, as determined by your quality metric, visualize the distribution of Autodock Vina scores in each cluster. Do members of the same cluster have similar docking scores? Why or why not?**

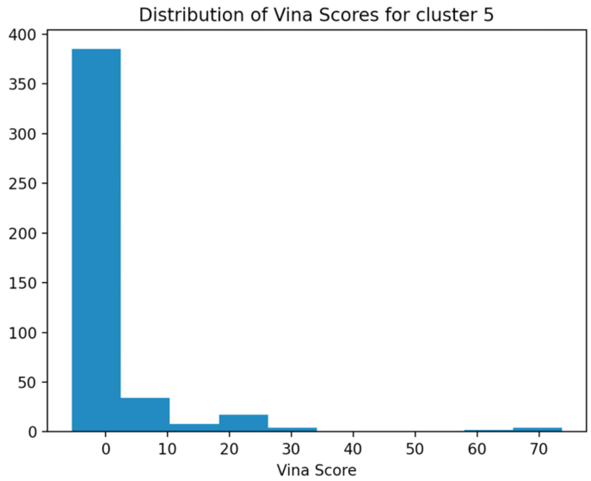
Partition clustering performed the best in question 7, so I will visualize the distribution of scores of groups generated with that clustering, with K = 5. I will also use the full dataset again.









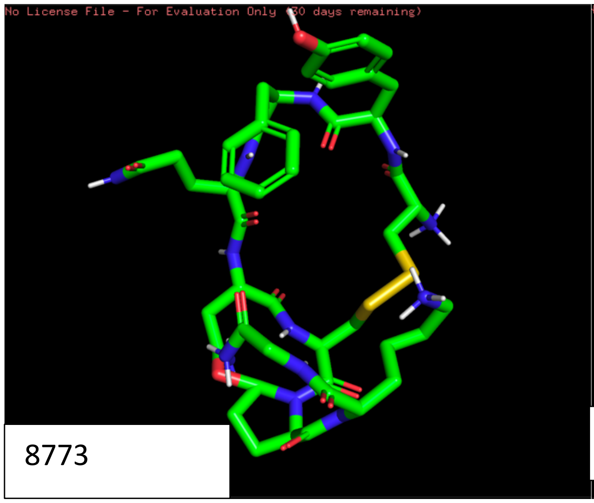


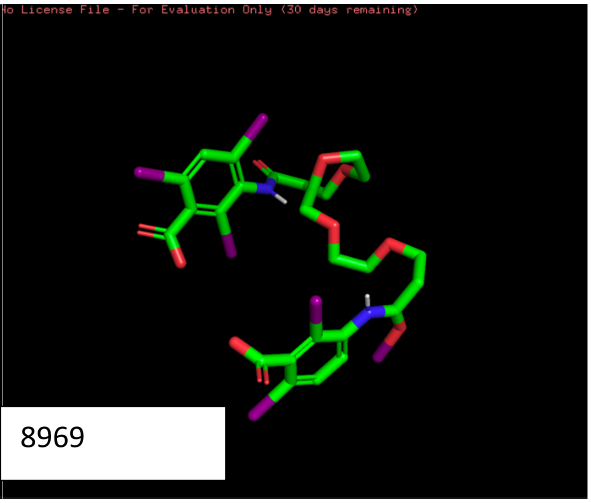
Yes, it does seem that Vina scores for each cluster are relatively close to each other. Most Vina scores are relatively close to zero, though it seems that cluster 2 has an enrichment of high-scoring ligands and cluster 3 has an enrichment of negative-scoring ligands (although this cluster is quite small in terms of population). The other clusters seem to have similar scoring ligands, most of them close to zero.

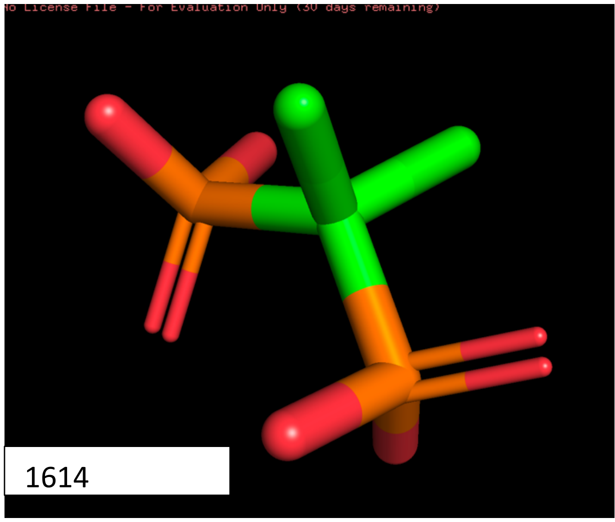
10. Select the top scoring molecule from each cluster. This is your list of cluster heads. Visualize the top 5 by score in PyMOL and pick your favorite. Are they structurally diverse?

Terminal output:

```
Highest scoring in cluster 1 is ligand 8773 with Vina score 107.7
Highest scoring in cluster 2 is ligand 8969 with Vina score 44.9
Highest scoring in cluster 3 is ligand 1614 with Vina score -3.2
Highest scoring in cluster 4 is ligand 8524 with Vina score 117.9
Highest scoring in cluster 5 is ligand 9055 with Vina score 73.0
```

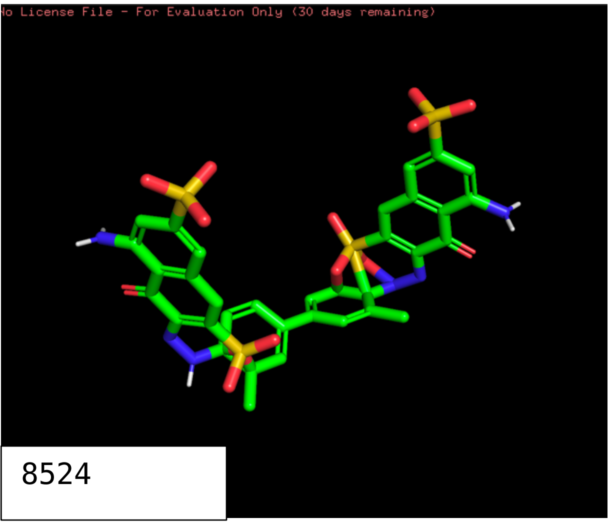






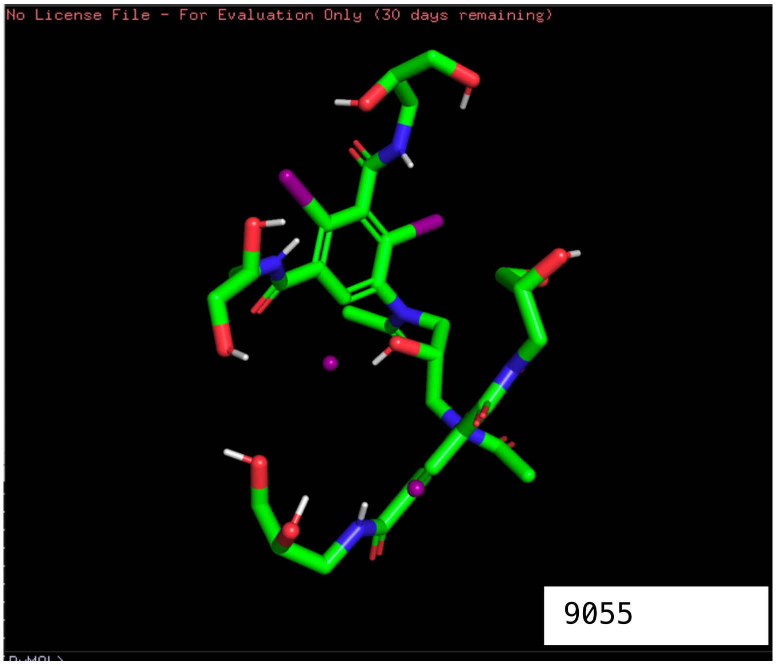
o License File - For Evaluation Only (30 days remaining)

1614



lo License File - For Evaluation Only (30 days remaining)

8524



No License File - For Evaluation Only (30 days remaining)

9055

Ligand 1614 is obviously distinct from the rest of them, both in terms of structure and size. The other 4 are somewhat visually similar in the sense that they seem to be long, chainlike molecules with at least one aromatic ring. My favorite is 8524 both because it is the highest scoring and it looks cool with all of its chained rings.