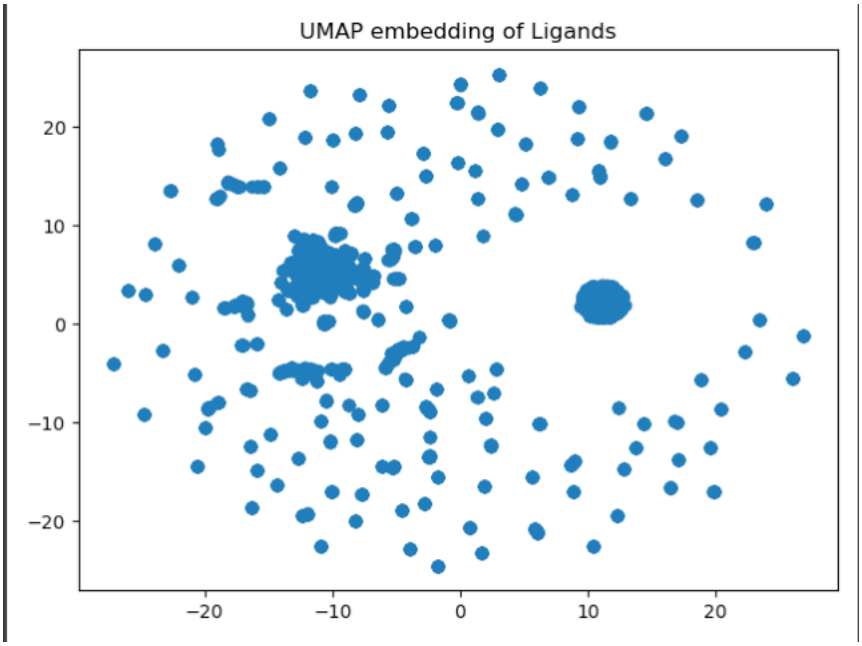
BMI 203 Winter Quarter 2021
Homework Assignment 2
Miguel Guardado
02/19/2021

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

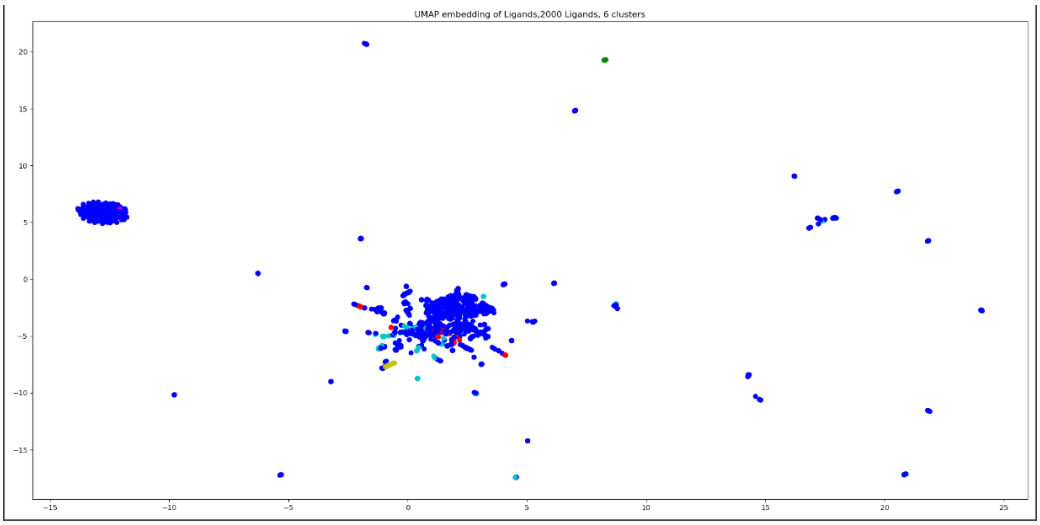
   The distance metric I chose to implement was a standard Euclidean distance. I chose this metric because of how standardizable is it to implement and use for data. Choosing to make a generalizable clustering algorithm that can be used for any type of data, not making it biologically/chemically focused, I wanted to create a prototypical distance metric that be utilized on any type of dataset. The way the algorithm was designed, can easily allow the extension of other distance metrics, since all distance calculations goes though the **CalculatePairWiseDistance(),** this can easily allow the extension were you input an extra parameter based off any distance metric you want to implement. I ran out of time, so I did not do this, but my code is extended to allow this if I desire.

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 will choose Umap as my dimension reductions algorithm, this will be implemented on all the ligands from the dataset. This implementation will be the raw data of each ligand, having the data be an array representation [0,1] of the 1024 dimension bit space.



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.
   For the rest of this analysis, this was ran on the first 2000 ligands, with duplicates left in there. Which is why some of the cluster visualizations are dominated by one color, since most of the clusters will encompass a cluster of repeated ligand ID's. Just a general note for the rest of the analysis.



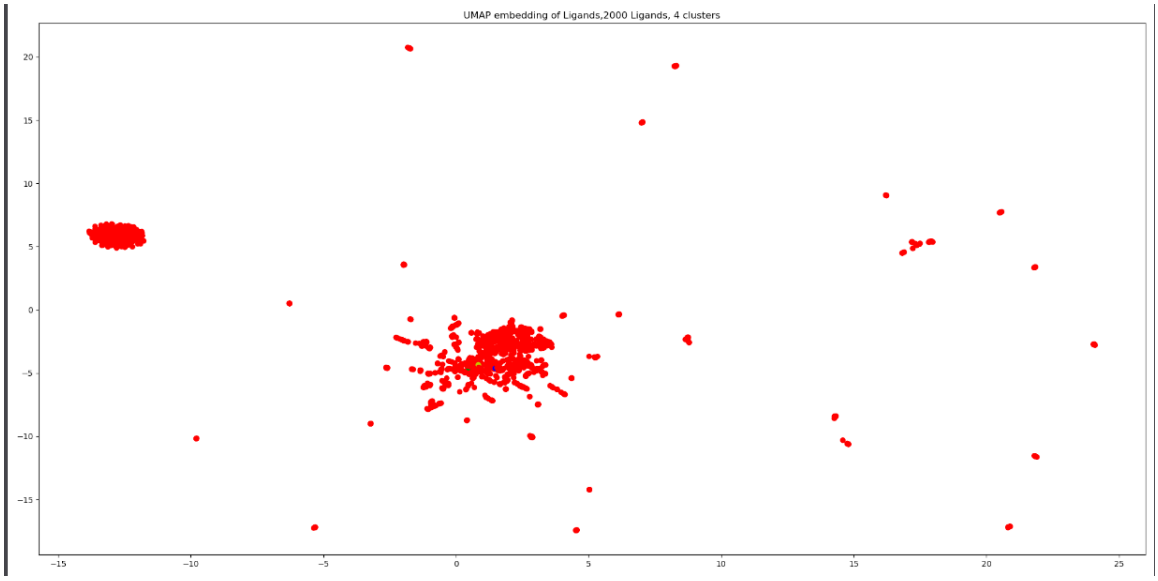
4. Explain your choice of partitioning clustering algorithm. Visualize this clustering in the same way as question 3

No. Clusters(k)	Sil. coeff
1	-1
2	0.2035
3	0.0933
4	0.18810
5	0.0485

No. Clusters(k)	Sil. coeff
6	0.396001
7	0.22705
8	0.08660
9	0.29346

I will use the silhouette coeff to determine the best fit, since it shows a measure of the quality of cluster created. The motivation is to use the highest score as my choice of clusters to use for my analysis. For Partition Clustering, I get the highest score at K=6 to use for my analysis. This means that K=6 for this data gives the best graphical representation of how well this algorithm was able to classify each cluster.

5. Cluster the small molecules using your implementation of a hierarchical clustering algorithm. Visualize this clustering in the same way a question 3.
This plot shown below is a visualization of



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

No. Clusters(k)	Sil. coeff
1	-1
2	0.0953
3	0.08660
4	0.26153
5	0.081803

No. Clusters(k)	Sil. coeff
6	0.163898
7	0.233848
8	0.09873
9	-0.167866

Similar to my response from question 4, I will use the silhouette coeff to determine the best fit. For Hierarchical Clustering, I get the highest score at K=4 to use for my analysis. This means that K=4 for this data gives the best graphical representation of how well this algorithm was able to classify each cluster.

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

   I will refer to my silhouette scores of each of the two tests that I ran to determine which clustering algorithm preformed the best. I chose to use a silhouette score as my choice of quality metric since this metric will look, for each individual observation, the average intra-cluster distance over the average inter-cluster distance, looking at the variance within a cluster vs the variance outside its own cluster. Along with the addition of the metric being scaled between [-1, 1], making it easier to compare to other clustering methods runs. Looking back at which cluster method preformed best between partitioning vs hierarchal clustering, I will compare the best score of each of the two algorithms together, K=[4,6]. Partitioning silhouette score is [4,6] is [0.18810, 0.396001] while Hierarchical is [0.26153, 0.163898]. The difference between these points (PT-HC) is [-0.07343, 0.2321] showing that the partitioning algorithm is out preforming the hierarchical at k=6 compared to the difference that it has at k=4. Doing a vector subtraction of each of the 9-clustering test run (PT-HC), keeping the distance a vector, not Euclidean distance, so we can utilize the sign of sum to see the average of what score preforms best. We get a value of 0.682669 showing that the partitioning algorithm is on average preforming better than the hierarchical clustering.

8. Compare the two-clustering using your implementation of a clustering quality metric. Explain your choice of quality metric. Which clustering preformed 'best' according to your metric

   When I calculate the Tanimoto coeff for K=4,6 I get a score of [0.898679,0.48718] respectively. With the higher value of 0.898679, this shows that using a k=4 clustering gets you a better measure of similarity, which is also shown by a shorter difference in distance compared to k=6(question 7). This leads me to the conclusion that the best cluster to use for this small section of 2000 ligands would be a k=4 clustering technique, with a preferred technique of k-means clustering due to it outperforming on average the hierarchical clustering technique.

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

   I apologize but I am unable to get to this section. But I put a lot of work into my algorithms + documentation so I am really happy with how it turned out compared to my previous project! It was really fun to learn sphinx, I didn’t realize how cool and effective it is to document code this way!! Something I want to continue to learn and use during my time in grad school! Thanks for pushing us to learn it, I think this is super useful to know!

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?

    Since I didn’t do 9, I also don’t know what my favorite score is, but I can share my favorite clustering meme I found over the course of this topic. I hope you enjoy my humor.