Journal of High School Science

Enhancement of drug discovery with machine learning clustering algorithms

Damarla R

Submitted: March 31, 2022, revised: version 1, April 19, 2022
Accepted: April 23, 2022

Abstract
Drug discovery is a process requiring lots of time and effort. However, for an infectious virus like COVID-19, this process is time-consuming. This study investigates potential ways to enhance drug discovery with machine learning clustering algorithms to cluster compounds based on the similarity of their chemical properties. The clustering algorithms that were chosen were K-Means and Hierarchical. K-Means was chosen under the assumption that there was a linear correlation between the compounds and their assigned clusters. Hierarchical was performed to understand the big picture of the dataset's clustering. The machine learning techniques successfully clustered 16816 compounds into roughly 100 different clusters. With even larger datasets and fine-tuning of the model future investigations could use the proposed clustering process to test a few compounds from each cluster rather than testing many more at random. This increase in the efficiency of drug discovery is based on the assumption that chemical compounds with similar physical properties, such as molecular formula, molecular weight, boiling point, melting point, density, and strength measured as pka, behave similarly against viruses and other foreign invaders. Thus, if a compound from a specific cluster exhibits an immune response then researchers can target that cluster and test the next closest compound in the cluster. This enhancement of the initial stages of the drug discovery process can help improve its speed and accuracy.

Keywords
COVID-19, K-Means clustering, Machine Learning, Dimensionality Reduction, Hierarchical clustering, ADMET optimization

---

[1]Corresponding author: Rishi Damarla, Academie Ste. Cecile International School, 925 Cousineau Rd, Windsor, Ontario, Canada N9G 1V8, 2017r_damarla@stececile.ca
Present Address: 1362 Tuscany Oaks Drive, LaSalle, Ontario, Canada N9J 0B6

Introduction

COVID-19 has infected more than 487 million people and killed over 6,000,000 as of March 31, 2022 (1). It has quickly become one of the deadliest viruses in modern history. Thus, scientists from around the world are working to find a cure for COVID-19. More specifically, there are two possible cures scientists are investigating - vaccines and antivirals. Antiviral development is quicker than that of vaccines, but it still requires a considerable amount of time and effort. However, due to the havoc that COVID-19 has caused, there is a sense of urgency to speed up the drug discovery process.

Past studies have tried to predict the outcome of clinical trials using machine learning algorithms (2). In 2019, Project ALPHA led by MIT Professor Andrew Lo used datasets featuring past clinical trials results to create a machine learning model to predict the probability of future trials' success (3). Although Lo's study didn't focus on drug discovery, it relied on past data to predict the future. However, with a novel virus, such as SARS-CoV-2, antivirals will have to be made to fight against the virus's DNA, which can only be sequenced after the virus's introduction. There are also several past studies conducted with the same aim in mind as this one such as a 2015 study that identified new candidate drugs for the treatment of lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm (4). This study aims to build on the aforementioned one by changing certain procedures. Instead of testing for possible antivirals immediately this study aims to categorize the compounds first and then test certain compounds from each cluster to determine the overall cluster's effectiveness instead of testing each compound individually.

Methods

Python

The programming language Python, version 3.7.6, was used to implement the following procedure (5) . Jupyter Notebook was the IDE used to code the algorithms (6). The following libraries were imported into Jupyter Notebook to design the clustering algorithms and create the visualizations: scikit-learn, pandas and numpy (7 - 9). The clustering algorithms used were K-Means and Hierarchical Clustering (10, 11).

Dataset Preparation

Clustering of Compounds

The dataset used to implement the clustering algorithms was found at the CAS Registry (12). This dataset includes 48876 compounds, some of which are known antivirals, and their chemical properties. The chemical features include the molecular formula, molecular weight, predicted and experimental boiling points, experimental melting point, predicted and experimental densities, and strength of the compound measured as the pka, which is a value that determines the acidity or basicity of a solution. The higher the pka value, the more basic a compound is. The chemical features that were selected for experimentation were chosen because they represent a diverse set of physical characteristics that can be used to divide the compounds into distinct clusters. However, melting point, boiling point, and density experimental were removed as features due to the missing values. This reduced the total number of compounds to 25738. Additionally, many of the remaining

compounds had molecular weights greater than 500g/mol which didn't satisfy the Lipinski Rule of five that outlines which compounds, based on their chemical properties, would have the best chance of being orally active in humans (13). One such property used in the Lipinski Rule is molecular weight, and it states that compounds with a molecular weight greater than 500g/mol are less likely to activate an immune response against an invading virus. Therefore, only compounds with a molecular weight of less than 500g/mol remained, which brought the number of compounds remaining to 16816.

Implementation of the Algorithms

Dimensionality Reduction with PCA, t-SNE, and UMAP

Before implementing the two clustering algorithms dimensionality reduction was performed on the data to reduce the number of initial features of the dataset to two or three principal features. Dimensionality reduction is necessary because it would be hard to visualize the results of clustering with a high number of features (14). The three dimensionality reduction methods used were Principal Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP).

The dimensionality reduction mechanisms work by projecting data points from the initial high dimensional space onto a new low dimensional space while maximizing the variance between the original data points (14).

K-Means Clustering

K-Means Clustering was one of the two clustering algorithms used on the CAS Dataset. K-Means works by separating data points into k clusters based on the distance between them, which is determined by their principal features. The number of k clusters used was 100 given the size of the dataset. The algorithm started by initializing 100 random points in the data as centroids, or cluster centers. Afterwards, the algorithm scrutinized each data point and the distance from each of the 100 centroids. Then it assigned each data point to its closest centroid and reevaluated the values of all the centroids by "summing up all the points belonging to that centroid and dividing the sum by the number of points in the group" (15). This process was repeated until the values of the centroids didn't change, and in the end, each cluster was separated based on the given principal features.

Hierarchical Clustering

Hierarchical clustering was the second clustering method used on the CAS Dataset. There are two types of hierarchical clustering, agglomerative and divisive. Agglomerative clustering is the bottom-up approach in which each data point is considered to be its own cluster. Divisive clustering is the top-down approach in which the entire pool of data is considered to be one cluster. For this study agglomerative clustering was used. After assigning each data point as a cluster, agglomerative clustering merged together the two clusters closest to each other based on the linkage function, a model hyperparameter that indicates the ways of calculating a distance between clusters. This repeated until all clusters were merged into one single cluster. The result of this clustering was a dendrogram or a tree diagram that illustrates all the possible ways of splitting the data into x number of clusters. After the clustering was applied to the

16816 compounds, three clusters were chosen based on the height of the dendrogram's y axis. These three clusters were plotted on a graph and then transformed into three different datasets that were used to perform further clustering on. The three new datasets were further clustered into 25-75 final clusters depending on their size.

Sampling Compounds

Once both clustering algorithms were performed on the compounds, random points from various clusters were taken as samples to verify that the models worked well to assign similar compounds to similar clusters and dissimilar compounds to dissimilar clusters. Random sample compounds were selected using the sampling method available in Python. This method returned the given number of sample compounds along with the values of their chemical features. This allowed for the observation of how compounds assigned to the same cluster differentiated from those assigned to different clusters.

Results

Visualization of K-Means Clustering

The three dimensionality reduction techniques were applied to the data before using the K-Means clustering algorithm. Figure 1 shows the visualization of the three different techniques with the raw data.

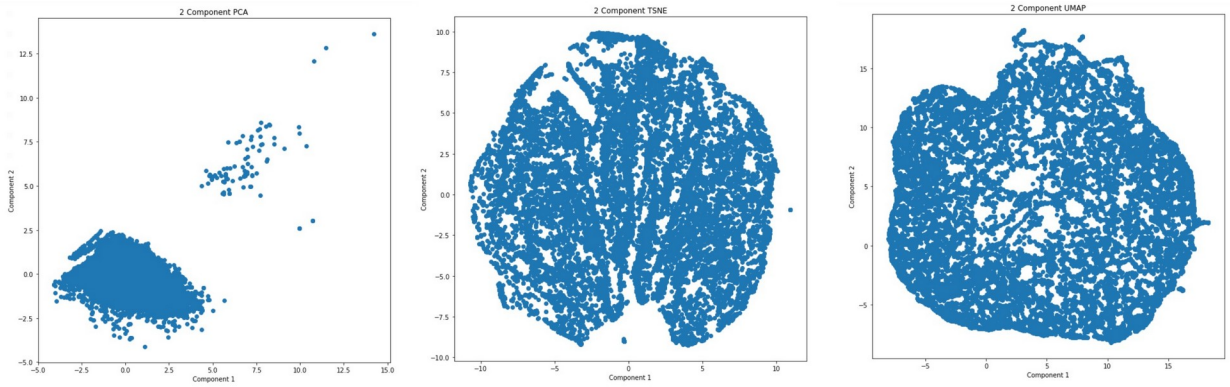

Figure 1: 2D Plots of PCA, t-SNE, and UMAP. Principal Component Analysis is shown in Panel A. T-distributed Stochastic Neighbor Embedding is shown in Panel B. Uniform Manifold Approximation and Projection is shown in Panel C.

Although the K-Means clustering was performed on each of the three panels shown in Figure 1, t-SNE (T-distributed Stochastic Neighbor Embedding) was chosen to be illustrated. Figure 2 shows the different steps of implementing the K-Means clustering on the t-SNE data that was previously shown in Panel B of Figure 1.

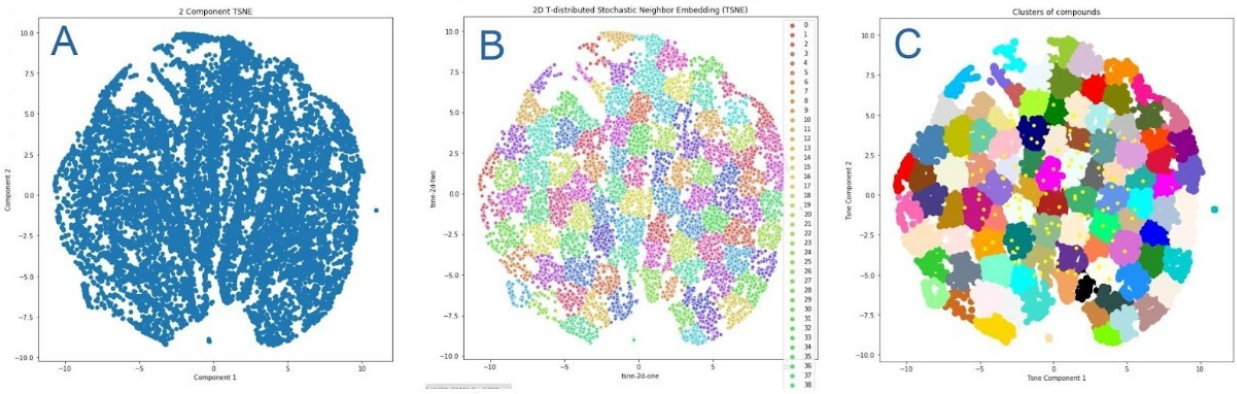

Figure 2: The different Stages of the 2D K-Means clustering process. The raw data before clustering are shown in Panel A, which shows that the dataset is diverse and well distributed over the chemical space. The data after clustering has been applied are shown on a scatter plot in Panel B. The final stage in which the data are visualized in the 100 clusters is shown in Panel C.

The previous figures show a 2D visualization of the clustering with only 2 principal features. However, the data can also be visualized on a 3D plane by adding an extra principal feature. Figure 3 illustrates the 3D plots of the data before applying the clustering algorithm.

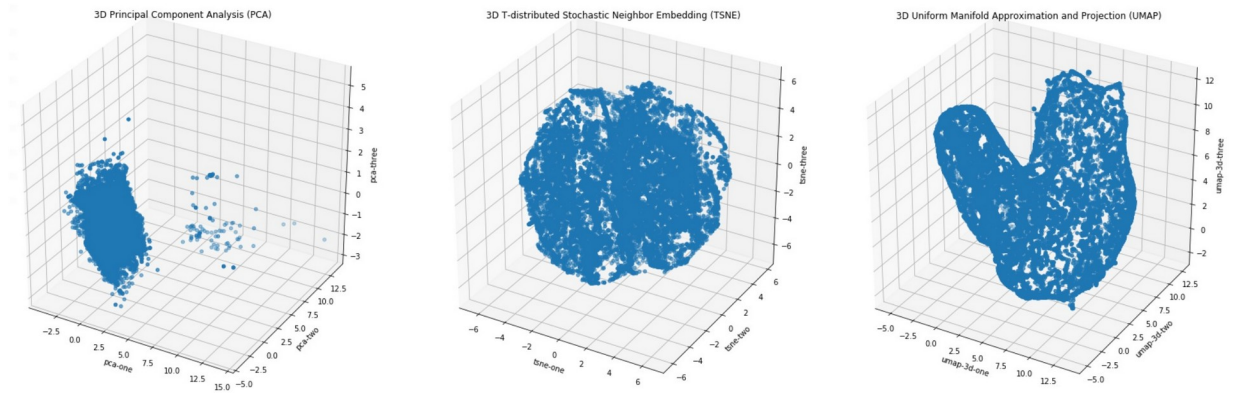

Figure 3: 3D Plots of PCA, t-SNE, and UMAP. Principal Component Analysis is shown in Panel A. T-distributed Stochastic Neighbor Embedding is shown in Panel B. Uniform Manifold Approximation and Projection is shown in Panel C.

Figure 4 shows the different stages of the 3D K-means clustering process.

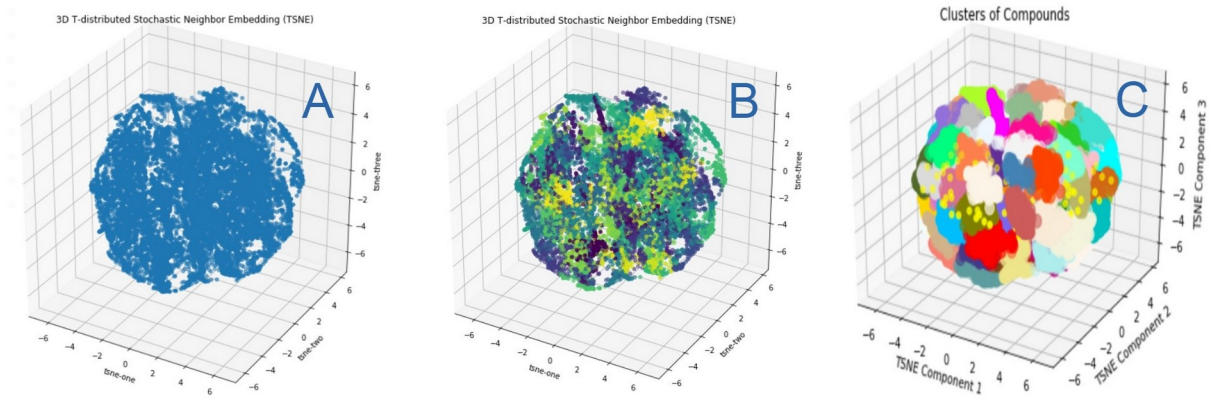

Figure 4: The different Stages of the 3D K-Means clustering process. The raw data before clustering are shown in Panel A. The data after clustering has been applied are shown on a scatter plot in Panel B. The final stage in which the data are visualized in the 100 clusters is shown in Panel C.

Sampling of Random Compounds from K-Means

Ten compounds were sampled randomly out of all 100 clusters to show that the model accurately clustered similar compounds together while separating dissimilar compounds. Figure 5 shows the 10 compounds and their chemical features as well as their principal features labelled as t-SNE-2d-one, t-SNE-2d-two, t-SNE-2d-one-3d, t-SNE-2d-two-3d, and t-SNE-2d-three-3d. The first two were used to plot the clusters on a 2D scale while the last three were used to plot them on a 3D scale. The figure also shows the clusters that each compound was assigned to. The column cluster1 refers to the cluster the compound was assigned to on a 2D plot while cluster2 refers to the cluster it was assigned to on a 3D plot.

CAS Index Name	Molecular Formula	Molecular Weight	Boiling Point Predicted	Density Predicted	Pka Predicted	tsne-2d-one	tsne-2d-two	cluster1	cluster2	tsne-2d-one-3d	tsne-2d-two-3d	tsne-2d-three-3d
Ethanimidamide, N'-(9-β-D-arabinofuranosyl-9H-...	C14H20N6O4	336.35	622.8	1.610	13.10	-5.994328	-2.252665	17	90	-3.103493	-3.056312	0.081070
Thymidine, α,α,α-trifluoro-3-[[(1-methylcyclop...	C17H21F3N2O7	422.35	502.6	1.540	13.89	-5.895068	1.256322	19	1	-2.104409	-0.514553	-0.571272
Furo[3',4':6,7]naphtho[2,3-d]-1,3-dioxol-6(5aH...	C20H16O7	368.34	589.5	1.524	13.32	-5.524950	-1.195855	91	90	-2.845542	-2.322762	-0.814258
4H-1-Benzopyran-4-one, 5-[(4-fluorophenyl)meth...	C20H16FN3O6	413.36	626.4	1.498	6.82	1.543007	-1.748276	98	88	1.963279	-0.828775	-0.566488
2H-3,1-Benzoxazin-2-one, 1,4-dihydro-6-methyl-...	C16H16F3NO2	311.30	324.9	1.260	11.41	-4.031394	7.987776	83	4	-2.650317	1.815301	4.846766
19,26,27-Trinorergostane-3,24-diol, 3-methyl-,...	C26H46O2	390.64	492.6	1.000	15.20	-5.250993	5.920400	32	17	-5.472275	1.810402	-3.875314
2(1H)-Pyrimidinethione, 4-(3,4-dimethoxyphenyl...	C22H28N2O2S	384.53	507.2	1.260	12.15	-2.400937	4.096788	80	99	-3.059411	1.463577	-1.930126
α-L-Xylofuranurononitrile, 1,2-dideoxy-1-(1,6-...	C11H10FN5O4	295.23	773.3	1.930	11.43	-4.070727	-6.837169	94	45	0.002828	-4.584702	1.532259
Guanosine, 2'-deoxy-N-(1-methyl-3-oxopropyl)-	C14H19N5O5	337.33	701.3	1.710	13.88	-4.300736	-4.693077	61	73	-1.793276	-5.473906	-0.125786
L-Alanine, N-(hydroxyphenoxyphosphinyl)-, meth...	C20H27N4O8P	482.42	643.4	1.510	13.92	-2.573836	-2.694079	67	82	-0.188296	-2.952677	-2.817499

Figure 5: Sampling of Random Compounds from K-Means. The chemical features of each of the ten compounds are shown along with the calculated 2D and 3D principal features and the clusters they were assigned to.

Visualization of Hierarchical Clustering

In addition to the K-Means clustering algorithm, hierarchical clustering was also performed on the CAS Dataset. The dimensionality reduction technique UMAP (Uniform Manifold Approximation and Projection) was chosen to be illustrated through the process of implementing the hierarchical clustering algorithm. Figure 6 shows the steps through which the unclustered data was separated into three clusters.

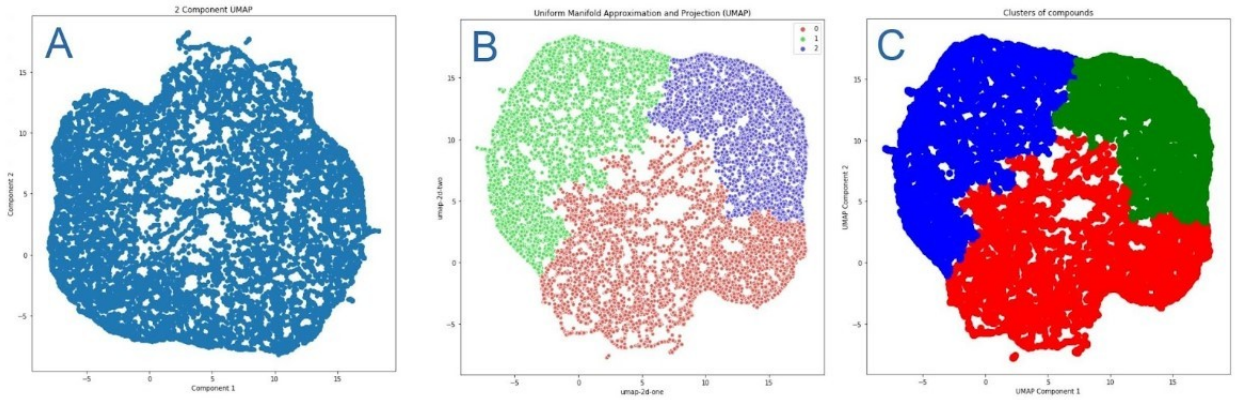

Figure 6: The different Stages of the 2D Hierarchical clustering process. The raw data before clustering are shown in Panel A. The data after clustering has been applied are shown on a scatter plot t in Panel B. The final stage in which the compounds are grouped together in 3 different clusters is shown in Panel C.

After the data was split into three large clusters, further clustering was applied to all three clusters. Figure 7 shows one of these three clusters in its process of undergoing further clustering.

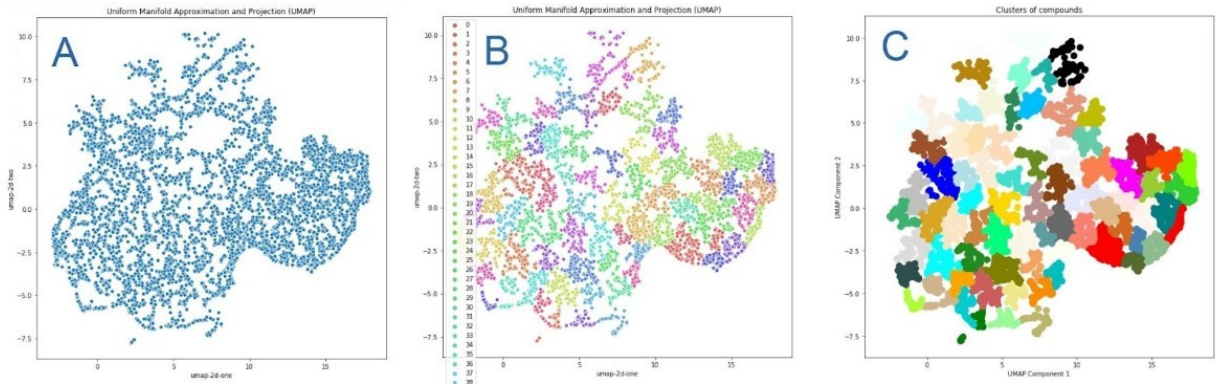

Figure 7: The different stages of Further 2D Hierarchical clustering process. The unclustered data of one of the previous clusters from Panel C of Figure 6 is shown in Panel A. The data after clustering has been applied are shown on a scatter plot in Panel B. The final stage in which the data are visualized in 75 clusters, because of the relatively smaller size of the data, is shown in Panel C.

The compounds were also visualized on 3D scatterplots with an extra principal feature added. Figure 8 displays the process of splitting the data into three large clusters on a 3D plane.

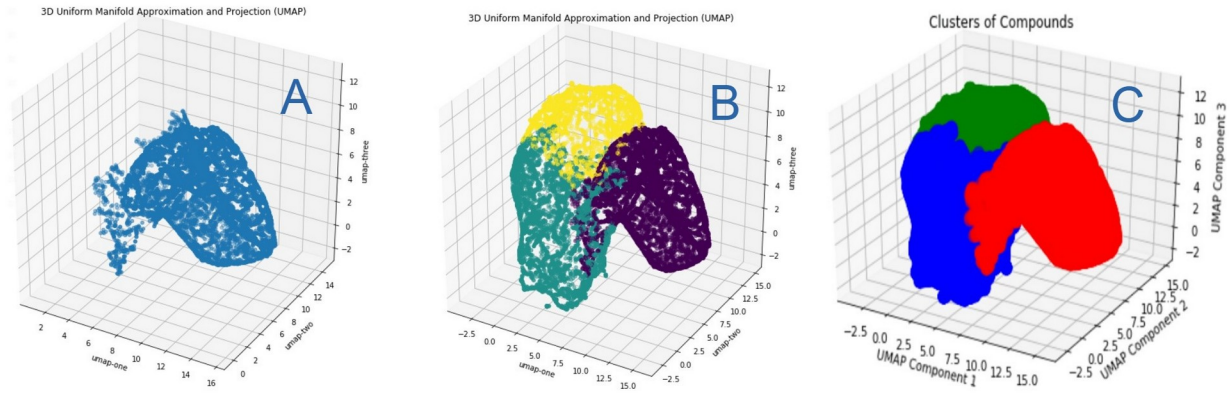

Figure 8: The different Stages of the 3D Hierarchical clustering process. The raw data before clustering are shown in Panel A. The data after clustering has been applied are shown on a scatter plot in Panel B. The final stage in which the data are visualized in the 3 clusters is shown in Panel C.

Lastly, the three clusters shown in Panel C of Figure 8 underwent further clustering. In Figure 9 one of the three clusters is visualized going through the three stages of the hierarchical clustering process.

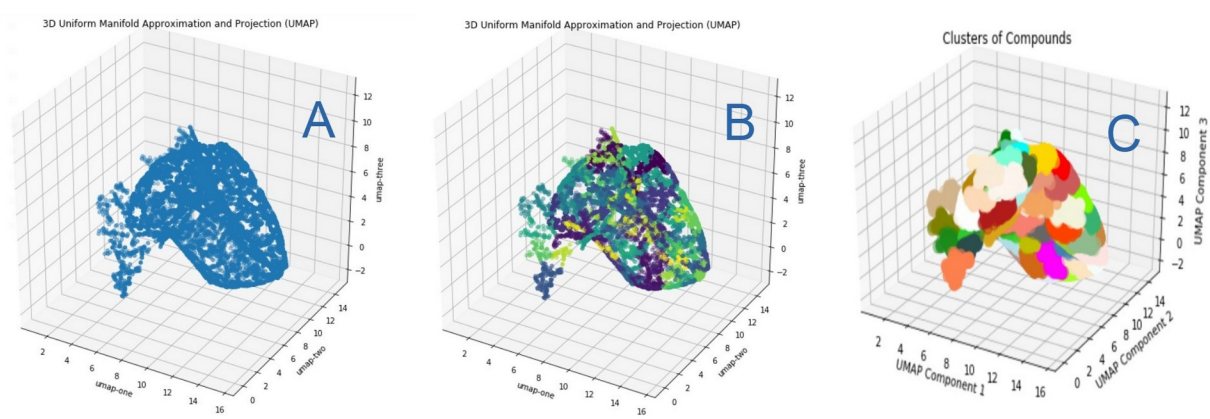

Figure 9: The different Stages of the 3D Hierarchical clustering process. The unclustered data of one of the three clusters are shown in Panel A. The data after clustering has been applied are shown on a scatter plot in Panel B. The final stage in which the data are visualized in the 100 clusters is shown in Panel C.

Sampling of Random Compounds from Hierarchical

Random compounds were also sampled from the hierarchical clustering model. The compounds clustered twice, first into three large clusters and then into 50-100 smaller clusters. This process is shown in Figure 10.

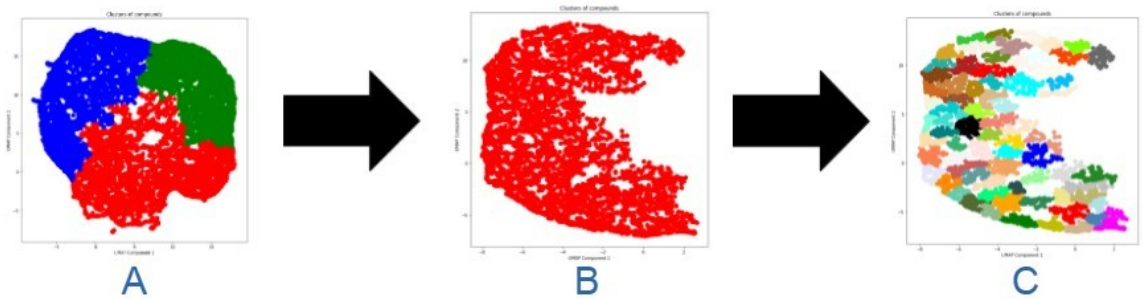

Figure 10: The different Stages of Further 3D Hierarchical clustering process. The data clustered into 3 clusters is shown in Panel A. The red cluster from Panel A is shown as a scatter plot in Panel B. The final stage in which the red cluster undergoes further clustering is shown in Panel C.

Figure 11 shows the random sampling of ten compounds from the hierarchical clustering process. The compounds were chosen from cluster 0 of the three clusters that were assigned at the beginning. Although these compounds were initially part of the same cluster, they were further clustered into 75 new clusters. The cluster when reclustered column shows the cluster that each compound was assigned to. The compounds were assigned to these clusters based on their principal features which are represented by umap-2d-one and umap-2d-two in Figure 11.

cluster	CAS RN	CAS Index Name	Molecular Formula	Molecular Weight	Boiling Point Predicted	Density Predicted	Pka Predicted	umap-2d-one	umap-2d-two	cluster	Cluster when Reclustered
0	874327-72-7	1H-Imidazo[4,5-c]quinolin-4-amine, 2-[(2-metho...	C22H24N4O2	376.45	604.1	1.250	6.08	-0.083546	-3.061313	0	54
0	1312899-57-2	4H-1-Benzopyran-4-one, 5-[(4-fluorophenyl)meth...	C26H20FN3O6	489.45	690.3	1.434	6.53	-3.953930	-5.624094	0	48
0	1987626-81-2	Benzonitrile, 2-[[2-[(2,6-dimethylphenyl)amino...	C19H17N5	315.37	535.8	1.250	4.93	-2.546276	0.618002	0	14
0	1626-66-0	9H-Purin-2-amine, 9-(5-O-phosphono-β-D-ribofur...	C10H14N5O7P	347.22	831.6	2.320	1.86	-3.795977	12.624568	0	47
0	147127-15-9	Phosphonic acid, P-[[(1R)-2-(2,6-diamino-9H-pu...	C9H15N6O5P	318.23	832.7	2.050	2.30	-2.569264	9.356544	0	50
0	1007389-94-7	9H-Purin-6-amine, 9-[2-[[4-[4-(1,1-dimethyleth...	C21H28N5O4P	445.45	646.6	1.400	4.21	-6.528055	-3.443637	0	28
0	374067-84-2	Benzonitrile, 4-[[4-(4-ethynyl-2,6-dimethylphe...	C21H16N4O	340.38	541.3	1.270	3.49	-7.817149	-0.578332	0	11
0	1037453-04-5	5-Pyrimidinecarboxylic acid, 4-[[4-[(1E)-2-cya...	C25H22N6O3	454.48	692.5	1.350	1.04	-5.893956	-1.389436	0	17
0	145431-62-5	Phenol, 2-[(7-chloro-4-quinolinyl)amino]-4,6-b...	C25H29ClN4O3	468.98	598.2	1.347	9.33	3.397529	-5.897861	0	68
0	869962-22-1	1,4,2-Dioxaphosphorin-3(2H)-one, 2-hydroxy-6-p...	C9H7O5P	226.12	389.1	1.530	0.08	0.935783	9.004811	0	60

Figure 11: Sampling of Random Compounds from Hierarchical. The chemical features of each of the ten compounds are shown along with the calculated 2D principal features and the clusters they were reassigned to.

Silhouette Scores of Clustering Algorithms

Although there is no way of verifying whether or not similar clusters were grouped together because of the lack of labelled data, silhouette scores can be used to determine mathematically if compounds with similar properties, using the principal components, were clustered together [16]. The formula for the silhouette score is (b - a) / max(a, b) where a is the mean intra-cluster distance and b is the mean nearest-cluster distance (16). Table 1 shows the silhouette scores of each of the algorithms in 2D and 3D.

Algorithms with Dimensionality	Silhouette Scores
2D K-Means	0.4092985
2D Hierarchical	0.34618327
3D K-Means	0.3530448
3D Hierarchical	0.35166672

Table 1: Silhouette Scores. The silhouette scores were calculated to see if the model was able to accurately classify compounds based on their chemical features.

The score is a number between -1 and 1. Values close to -1 suggest that the model didn't have enough data to differentiate between the given data points whereas values close to 1 indicate that the model was able to successfully cluster most to all data points. The scores achieved for the algorithms, shown in Table 1, are between the values 0 and 1. These scores suggest that the clustering algorithms were able to create distinct clusters; however, the scores were not close enough to 1 to conclude that they created the most optimal clusters. These results could have occurred due to a suboptimal number of clusters that were used as parameters for K-Means and Hierarchical.

Discussion

This study aimed to determine whether or not machine learning algorithms could possibly be used to cluster compounds together thus improving the efficiency of the lengthy drug discovery process. Looking back at Figure 6 and Figure 8, one can see that hierarchical clustering accurately created clusters with certified boundaries despite arriving at a slightly different result compared to K-Means due to the differences in the algorithm's calculations.

The two algorithms clustered compounds based on their chemical properties. These properties are effective features as they can determine whether or not a group of drugs are chemically similar. Both of these clustering algorithms, K-Means and Hierarchical, achieved a good separation amongst the different clusters. In K-Means clustering the majority of the 100 clusters were distinct without much overlap, indicated both in the 2D and 3D visuals after clustering and the properties of sampled compounds in Table 1. In the Hierarchical clustering the compounds were first split into three large clusters contrary to the 100 initial clusters formed from K-Means. Afterwards the three large clusters with thousands of

compounds in each were divided into a number of smaller clusters depending on their size. In other words, Hierarchical clustering was implemented twice, generating two sequential models.

Although this study produced vivid results that prove that this method could be applied to speed up the process of drug discovery, there are certainly some limitations that need to be addressed. One such limitation includes the size of the dataset. Even though roughly 16000 compounds were tested after the initial screening steps, there are hundreds of thousands of compounds that scientists could look into during the exploration phase of drug discovery (17). However, due to the increased computational power that is widely available nowadays, there should not be any problem trying to replicate the clustering mechanisms with larger datasets (18). The only step that would have to change would be the number of random centroids initialized before running the algorithm. Another cause of concern is the number of chemical features used to perform the K-Means and Hierarchical clustering. In this study only 4 features, molecular weight, boiling point, density, and pka, were used to cluster the compounds. However, features can always be added or removed depending on their effects on the distribution of the clusters.

This research is only the beginning of how machine learning can help speed up the lengthy drug discovery process (18). In fact, not only could machine learning help with efficiency it could also help improve the accuracy of finding effective drugs (19).

Conclusion

The main goal of this study was to utilize big data and machine learning algorithms to develop workflows to speed up the drug discovery process. This process takes a considerable amount of time but with an infectious virus like COVID-19, this process is inefficient. Since the emergence of big data and machine learning it was hypothesized that machines could perform the task of exploring available drug compounds much more efficiently than humans (20). In this work, K-Means and Hierarchical clustering algorithms were implemented to organize compounds into separate groups based on their chemical properties. This model could be applied to allow researchers from different institutions and pharmaceutical companies to test a few compounds from each organized cluster instead of testing many more at random. This research provides an innovative solution, that can be implemented instantly, to speed up the drug discovery process.

References

1. "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)", coronavirus.jhu.edu/map.html.

2. Kent, J. "Machine Learning Tools Help Predict Clinical Trial Outcomes", 8 July 2019, healthitanalytics.com/news/machine-learning-tools-help-predict-clinical-trial-outcomes.

3. Baskin, K. “Using machine learning to better predict clinical trial outcomes”, 2 July 2019, mitsloan.mit.edu/ideas-made-to-matter/using-machine-learning-to-better-predict-clinical-trial-outcomes.

4. “Identification of new candidate drugs for lung cancer using chemical–chemical interactions, chemical–protein interactions and a K-means clustering algorithm.”, Taylor and Francis Online, www.tandfonline.com/doi/abs/10.1080/07391102.2015.1060161.

5. “Welcome to Python.org.” Python.org, www.python.org/.

6. “Project Jupyter.” Jupyter, jupyter.org/.

7. “Pandas.” Pandas, pandas.pydata.org/.

8. “NumPy.” NumPy, numpy.org/.

9. “Scikit-Learn Machine Learning in Python.” Scikit-learn, scikit-learn.org/stable/.

10. “Sklearn.cluster.KMeans.”, Scikit-learn, scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

11. “Plot Hierarchical Clustering Dendrogram.” Scikit-learn, scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html.

12. “Download CAS COVID-19 Antiviral Candidate Compounds Dataset.”, CAS Registry, www.cas.org/covid-19-antiviral-compounds-dataset.

13. Munir, A., Elahi, S., & Masood, N. (2018). Clustering based drug-drug interaction networks for possible repositioning of drugs against EGFR mutations: Clustering based DDI networks for EGFR mutations. Computational biology and chemistry, 75, 24-31.

14. Sarveniazi, A. (2013, December 15). An Actual Survey of Dimensionality Reduction. American Journal of Computational Mathematics, 2014, 4, 55-72.

15. Al-Masri, A. “How does k-Means Clustering in Machine Learning Work.”, Towards Data Science, 14 May 2019, towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0

16. “sklearn.metrics.silhouette_score.”, Scikit-learn, scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html.

17. “Stages of Drug Development.” Pacific Biolabs, pacificbiolabs.com/stages-of-drug-development.

18. “Large Datasets.” Science Direct, www.sciencedirect.com/topics/computer-science/large-datasets.

19. “Machine learning approaches to drug response prediction: challenges and recent progress.” Nature, https://www.nature.com/articles/s41698-020-0122-1.

20. “How artificial intelligence is changing drug discovery.” Nature, www.nature.com/articles/d41586-018-05267-x.