Graded out of 30 points

Adjusted grade (out of 15): 14/15

Good job describing the tasks, your solutions, and the reasons behind your design choices.

Exploring potentially effective chemical for COVID-19: similarity-based network analysis

Xingyu Lu
luxingyu@umich.edu
Data Science

Yueyang Shen
petersyy@umich.edu
Data Science

1 INTRODUCTION

In the past few weeks, the epidemic of the novel coronavirus, known as COVID-19, has been spreading rapidly across Europe, North America, Latin America, and Middle East. By April 20th, the number of cases across the globe surpass 2.3 million according to WHO[5],[9]. The current antiviral guidelines proposed by National Health Commission (NHC) of People's Republic of China include Chloroquine phosphate, IFN-α, lopinavir and several antiviral chemical for Prevention, Diagnosis, and Treatment[6]. While many clinical trials on drugs like Remdesivir, Hydroxychloroquine and Chloroquine are well under clinical trial phases, there are no drugs nor other therapeutics approved by the US Food and chemical Administration to prevent or treat COVID-19[7]. In light of this, we will be taking a computational approach to exploit the chemical space and disease space to better understand their associations to interpret the underlying structure and pattern in such spaces.

In this project, we mainly applied similarity-based unsupervised learning method in data-mining to uncover the underlying patterns within these spaces. Specifically, we applied **Louvain Algorithm**, **Spectral Clustering** and **K-means**, and we make use of some essential concepts from this course, such as Distance and Jaccard Similarity. Our inferences and predictions will be compared with the current knowledge and evidence on the disease.

We use 2 no-penalty late days for both team members.

2 DATA

We leveraged different datasets in the Comparative Toxogenomics Database (CTD)[4], which provides association between chemicals, phenotype, genes, disease, and pathways. We mainly utilized two datasets: chemical and disease association dataset, and disease and phenotype association dataset. The chemical and disease association dataset contains 1048547 entries of associations (March. 29^{th} version, 107MB). The other dataset, phenotype and disease association dataset also contains 1048547 entries (March. 29^{th} version, 31MB). Both datasets contain essential information related to COVID-19. In the chemical and disease association dataset there is a total of 1579 entries that is available and related to COVID-19. In the phenotype and disease dataset, we have a total of 1057 entries that are related to COVID-19.

3 DATA ANALYSIS

3.1 Q1: What are the common properties and underlying structure of the chemicals that are associated with COVID-19?

3.1.1 Data. We utilize the chemical and disease association database available on the CTD website. We select 1579 entries that are associated to COVID-19, and based on that we obtain 640 unique chemicals that are related to COVID-19. 515 of them have valid SMILES encoding, while 125 of them are not transferrable to SMILES in *cirpy*. We then transform these valid SMILES encoding to chemical fingerprints[3] where we are able to compute Tanimoto coefficient on. Among the chemical that have direct evidence on *COVID-19, emodin, acetaminophen, and (2-tert-butoxy-1-(2-cyclohexyl-1-(1-formyl-2-(2-oxopyrrolidin-3-yl)ethylcarbamoyl)ethylcarbamoyl) propyl) carbamic acid benzyl ester* are included in the 515 chemical set. We also note that compounds like Active Hexose compound, which has directed evidence in treating coronavirus[8], are among those 125 chemicals that are not included as we aren't able to transfer them to SMILES encoding, therefore we are unable to perform further manipulation on such chemicals.

It is also important to note that the original chemical disease association dataset did not include popular chemical and drug compounds like remdesvir and losartan, which might be proven effective for treating coronavirus in humans. Therefore, as we have not come up with a proper way to integrate all of them into our compiled dataset, they are not included in the upcoming discussion.

3.1.2 Technique & Challenges. A great challenge is to find a proper workflow for measuring the similarity between two chemicals given their name. Our finalized strategy is to transform it into SMILES and then to chemical fingerprints (Fig.1). We utilize a common similarity metric in chemoinformatics[12]. The Tanimoto coefficient[11], which is essentially the Jaccard similarity[14]

$$T_c = \frac{\text{number of } (1,1) \text{ rows}}{\text{number of non } (0,0) \text{ rows}} \quad (2)$$

The coefficient makes computing the similarity between two COVID-19-related chemical feasible. Naturally, we have obtained a fully-connected weighted graph G(V,E), where V consists of 515 candidate chemicals and E consists of similarity measure between any two nodes. In order to make the data more interpretable and visualizable, we transform this weighted graph into an unweighted graph, where we introduce a cutoff value β. That is, we would disconnect two nodes whenever the similarity between them is smaller than the cutoff value[13].

The reason why we choose to transform this graph instead of applying spectral clustering on the original weighted graph is that we want a more intuitive visualization for the underlying structure of the connections. Therefore, while Louvain is more computationally expensive, it is conceptually more straightforward than spectral clustering and since we are only dealing with a rather small amount of chemical (515), we do not have to worry about our computation time too much.

3.1.3 Experimental Setup. Our workflow for this part is the following

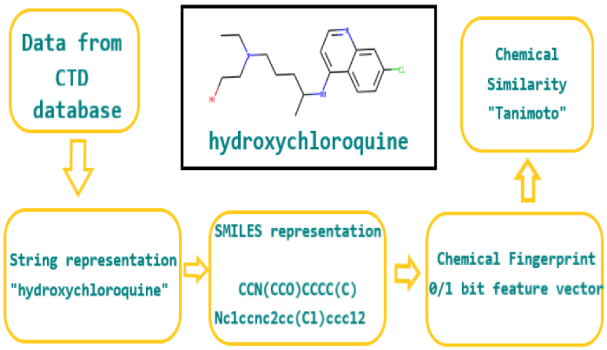

Fig.1 An example workflow of how chemicals are processed

3.1.4 Observations. We generate two undirected unweighted graph of the candidate chemicals in *gephi* (A clearer graph is attached at the end). The underlying "closeness" and "connectivity" of the underlying chemical space can be visualized as follows

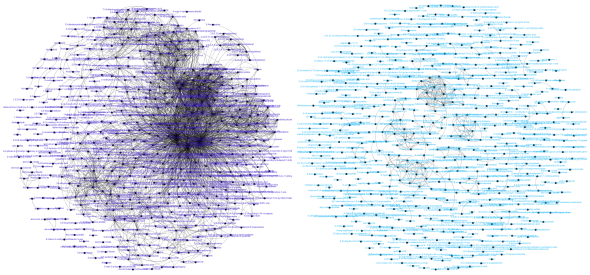

Fig.2 Chemical-Chemical Similarity Using Tanimoto Coefficient With Cutoff T_c >0.5(left) and T_c>0.7[13](right)

We note that the modularity on the right is 0.863, while the modularity on the left is 0.503. Both of them means that there is a community structure within the drug clusters. We then apply Louvain algorithm to detect the communities within the candidate chemical set.

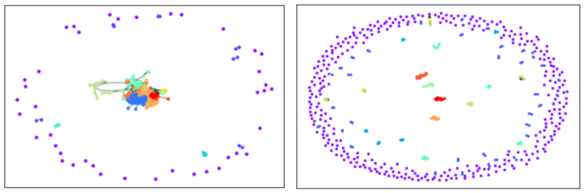

Fig.3 Chemical Community Detection Using Louvain algorithm for Dense graph T_c >0.5(left) and Sparse Graph T_c>0.7(right)

We note that the Louvain algorithm result parallels the visualization in gelphi, as on the right part, we observed sparse communities while on the left part we have relative dense communities (clusters). One of the dominant communities in the lower right graph is the **adenosine family**. They have a total of 18 members in their community and a visualization of these 18 members is provided in the data_preprocessing_drug notebook. As we aren't professionals in pharmacology, we aren't in a good position in evaluating the efficacy of such drug clusters instead we postulate two proposals:

- The outliers and "sparsely connected" neighborhoods could have potentially useful drugs that have been neglected.
- For a specific drug with direct evidence in treating COVID-19, say, acetaminophen[8], it is also important to take a closer look at its "neighbors" in the chemical space in terms of its potential application in treating COVID-19.

3.2 Q2: What are the diseases that are associated with COVID-19?

3.2.1 Data. For this part, we utilize the phenotype and disease association database from the CTD website. The database contains 13439 different phenotypes and 6033 different disease. A sample row of the database is as follows:

⟨phenotype name, phenotype ID, disease name, DiseaseID,...⟩

We extract all of the information needed from the first two columns.

3.2.2 Technique & Challenges. In this part, we mainly use spectral clustering to detect community that contains COVID-19. To do this, we first need to construct an similarity matrix representing the similarities between each pair of disease, and then apply spectral clustering. The similarity measure here is Jaccard similarity as well.

$$sim(\text{disease}_1, \text{disease}_2) = \frac{|\text{phenotypes}_1 \cap \text{phenotypes}_2|}{|\text{phenotypes}_1 \cup \text{phenotypes}_2|} \quad (2)$$

could have used LSH

The dataset has 6033 disease and 13439 phenotypes. Computing all the similarities is time consuming(it takes about 10 seconds to calculate similarities between diseases). To reduce the running time, we decide to first filter disease based on similarity against COVID-19, and do spectral clustering on the reduced list of diseases.
We notice that the similarity matrix is not very reflective since we only measure the similarity based on phenotype of two disease. For example, the similarity between COVID-19 and SARS is only 0.2904, while the similarity between COVID-19 and burns is 0.4151. This means we have neglected some latent factors. Therefore, we use this similarity measure as a preprocessing stage to filter out unrelated disease. Then, we decide to use Gaussian distance between similarity vectors of diseases to incorporate this latent information. We then apply spectral clustering on the weighted graph constructed from the Gaussian distance[10]. The weights are given by

$$w_{ij} = e^{-\frac{||s_i - s_j||^2}{\gamma}}$$

where s_i is the similarity vector consists of similarities between disease i and other disease. For simplicity we set γ to 1.

3.2.3 Experimental Setup. Work flow for this part is as follows:

3.2.4 Observations.

(1) **<COVID-19, other diesase> similarity**
The histogram of phenotype-base Jaccard similarity of COVID-19 and other disease is as follows: From the histogram, we can find that there are about 3400 disease that have few common phenotype with COVID-19. We set a threshold to filter disease with low similarity with COVID-19. Top-3 disease and some related disease are as follows: From the table, the similarity is not very reflective. So we decide to apply Gaussian distance between similarity vectors of disease.

-1 some type of evaluation?

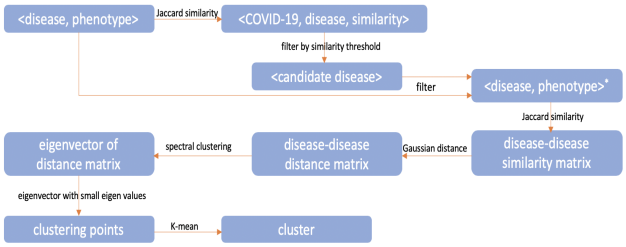

Fig. 4 work flow for Q2

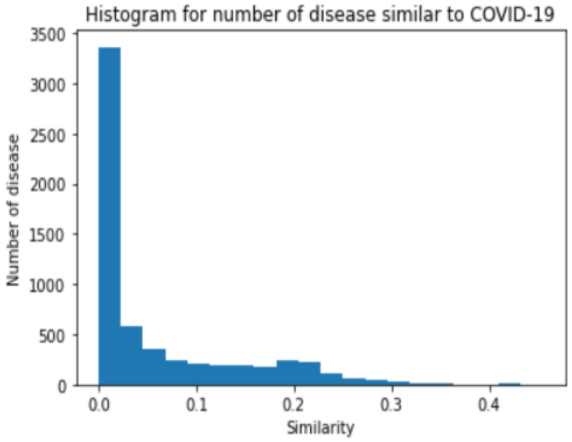

Fig. 5 Histogram of disease-COVID19 similarity

disease	similarity against COVID-19
Pleurisy	0.4545
Leishmaniasis	0.4290
Heat Stroke	0.4185
SARS	0.2904
Fever	0.2790
Pneumonia	0.2780

(2) **Gaussian Distance**
After applying the Gaussian distance measure, the three disease that are closest to COVID-19, and some related disease are as follows: The mean distance is 8.546×10^{-2}, median is

disease	distance from COVID-19
Pulmonary Edema	2.087×10^{-4}
Vomiting	4.303×10^{-4}
Nausea	4.700×10^{-4}
SARS	1.896×10^{-3}
Fever	1.256×10^{-3}
Pneumonia	4.470×10^{-3}

3.982×10^{-2}. So we can say that Gaussian distance is more reflective in detecting the community of COVID-19.

(3) **Clustering**

We construct the weighted graph based on the gaussian distance matrix. The visualization of eigenvectors and the TSNE plot of the clusters are as follows(we choose k = 5):

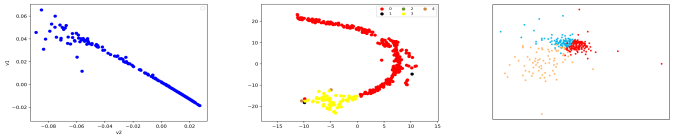

Fig. 6 Left: eigenvectors, Middle: tSNE plot, Right: Louvain(k=3)

In the Louvain plot, we omitted the outliers. In the tSNE plot, the points from same cluster are close to each other. Nevertheless, from the output of the algorithm, COVID-19, Fever, SARS are all in cluster 0. So we decide to use this clustering result for the next part as its result makes more intuitive sense.
Side note: There are 116 diseases in the cluster that COVID-19 belongs to (cluster 0).

3.3 Q3: What are the chemicals that are associated with the disease cluster of COVID-19?

3.3.1 Data. In this part, we use the chemical and disease association database and the results from Q1 (>0.5) and Q2: **1.** a csv file containing chemical and chemical cluster ID **2.** a csv file containing disease and diseaseID. We use MESH(Medical Subject Heading) ID of disease chemical to represent a certain disease or chemical.

3.3.2 Technique & Challenges. The main challenge in this part is to visualize the relationship between the chemical clusters and the disease clusters and we mainly depend on the https://graphonline.ru/en/ website that can visualize a weighted graph given it's adjancency matrix.

3.3.3 Experimental Setup. The work flow is as follows:

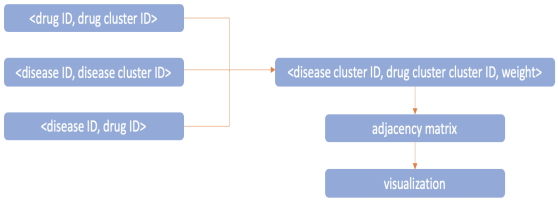

Fig. 7 work flow for Q3

3.3.4 Observations. With the adjacency matrix, we plot the weigthed graph with the help of https://graphonline.ru/en/ website. And the results are as follows:

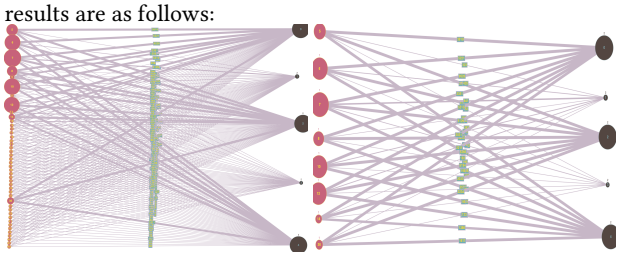

Fig. 8 chemical disease cluster relationship

The left graph is the graph of all the clusters obtained from question 1 and question 2. Nodes representing chemical clusters are aligned at left, and disease cluster nodes are on the right. There are node index on the nodes, disease clusters are 0-4 and chemical clusters are 5-48. The size of a node represents the importance of that node, the larger a node is, the more important it is. So we remove the chemical cluster nodes with low importance(these nodes also have few connection with the disease cluster nodes), and obtained the right graph.
Also, we plot the subgraph that only contain chemical clusters and disease cluster 0:

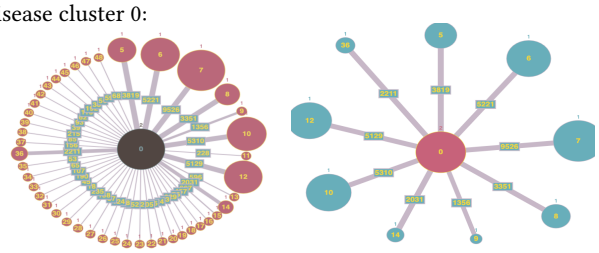

Fig. 9 chemical clusters COVID−19 disease cluster relation
The node 0 is the disease cluster that contains COVID-19. And if we set the weight threshold to 1000, we get the right graph, representing the 9 chemical clusters that are most related to the disease that similar to COVID-19. All the chemical in the 9 clusters are as attached in the appendix.
We count the number of chemical in each of the clusters as follows:

Node ID	5	6	7	8	9	10	12	14	36
number of chemical	42	37	84	20	11	45	32	6	3

Then we take both the weight and number of chemical in the cluster into consideration, we find that the node 36 contains chemical that are most closely connected to the community that contains COVID-19. And these chemical are aluminum ammonium sulfate(MESHid: C059726), aluminum sulfate(MESHid: C041524) and Ammonium Sulfate(MESHid: D000645). This result coincide with the data in CTD database as aluminum compound has second highest inference score with COVID-19[8].

4 CONCLUSIONS & DISCUSSION

Discuss the following:

(1) **Key Observations:**
1. Adenosine family is the most dominant chemical family that is associated with COVID-19.
2. Both drug graph (Fig.2) demonstrates community structure.
3. We observed several chemical clusters that are closely related to the disease cluster that contains COVID-19. And the chemical in the cluster that is most closely related to COVID-19 cluster are aluminum ammonium sulfate, aluminum sulfate and ammonium sulfate.

(2) **Challenges:**
- Real life is hard. There are so many design restrictions that we omitted like toxicology and side-effects and the details of a specific interaction (suppress or activate).
- Find a proper metric to measure the similarity of chemicals and disease, and how to visualize the structure in these spaces.
- Balance the tradeoff between finding a proper general dataset to manipulate on and being able to derive useful information from it.
- Come up with a controllable topic.
- Incorporate other's ideas into a brand new scenario carefully, thinking what would work, what won't.
- The barrier of trying to derive useful information as non domain experts.

(3) **What did you learn by doing this project?**
1. To begin with, we have learnt interesting ideas in evaluating and representing chemicals in the realm of chemoinformatics and bioinformatics. We really appreciate the brilliant idea of capturing chemical properties as chemical fingerprints and SMILES encodings.
2. The quality of source data as well as the data cleaning process determines the performance of the whole project, and the understanding of the source data is very important as well. In our database, there's a """ symbol in some entries. We first ignore this, and later, we found that this symbol comes a "," in the drug name entry. This cause much trouble as the whole database use "," as deliminator, and we have to rerun our whole jupyter notebook.
3. Data pre-processing is important. In big data analysis, running time is extremely important. When we generate the similarity matrix of different disease, we first run the program on the source data, and it runs for about 1 hour and didn't finish. And we didn't add any indicator to show the process of the program, so I interrupted the program. Then I tested one iteration, it takes for about 10 seconds, and the whole program requires 6000 iteration. That is, it would take about 10 hours to run the program on the whole database, while applying data pre-processing technique could be a huge savior. Also, this reminds me that some indicators are important for programs that takes a long time, such as printing the iterator so that we can know whether the program is running well.
4. Of course, we have a more thorough understanding on community detection and graph after this project. We learnt interesting ways to visualize a graph, both weighted and unweighted, directed and undirected.

(4) **Highlight:** We like the first part most. In this part, we learnt to use the "Gephi" software to visualize the graphs. We are both impressed by the elegant representation of the graphs in Gephi. They are attached in the appendix.

(5) **Contribution:**
Yueyang:
Brain storming the topic, find the database, literature review, writing Q1, programming Q1, visualizing some graphs
Xingyu:
Brain storming the working flow, writing Q2 and Q3, programming Q2, editing the report
We agree that each person has **50%** contribution to the project.

5 CITATION

(1) Zhang, Wen et al. “Predicting chemical-disease associations by using similarity constrained matrix factorization.” BMC bioinformatics vol. 19,1 233. 19 Jun. 2018, doi:10.1186/s12859-018-2220-4

(2) Wang, L et al. “Systematic analysis of new chemical indications by chemical-gene-disease coherent subnetworks.” CPT: pharmacometrics systems pharmacology vol. 3,11 e146. 12 Nov. 2014, doi:10.1038/psp.2014.44

(3) RDKit: Open-source cheminformatics; http://www.rdkit.org

(4) Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, Carolyn J Mattingly, The Comparative Toxicogenomics Database: update 2019, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D948–D954, https://doi.org/10.1093/nar/gky868

(5) “Coronavirus Disease (COVID-19) Outbreak Situation.” World Health Organization, World Health Organization, www.who.int/emergencies/diseases/novel-coronavirus-2019.

(6) "Guidelines for the Prevention, Diagnosis, and Treatment of Novel Coronavirus-induced Pneumonia", The 7th ed. http://kjfy.meetingchina.org/msite/news/show/cn/3337.html

(7) "Information for Clinicians on Investigational Therapeutics for Patients with COVID-19"(April 13th version), https://www.cdc.gov/coronavirus/2019-ncov/hcp/therapeutic-options.html

(8) CTD database COVID-19 page. http://ctdbase.org/detail.go?type=diseaseacc=MESH%3aC000657245view=chem

(9) Kupferschmidt, Kai et al. "Race to find COVID-19 treatments accelerates." Science vol. 367,6485. 27 Mar.2020, http://science.sciencemag.org/content/367/6485/1412.abstract

(10) Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395-416.

(11) Lo, Ben Torres, Jorge. (2016). Chemical Similarity Networks for Drug Discovery. 10.5772/65106.

(12) Rácz, A., Bajusz, D. Héberger, K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. J Cheminform 10, 48 (2018). https://doi.org/10.1186/s13321-018-0302-y

(13) Lo, Y.C., Senese, S., Li, C.M., Hu, Q., Huang, Y., Damoiseaux, R., Torres, J.Z. "Large-scale Chemical Similarity Networks for Drug Target Profiling of Compounds Identified in Cell-based Chemical Screens." PLoS Comput Biol. 11(3) (2015) [PMID:25826798]

(14) Bajusz, Dávid et al. “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?.” Journal of cheminformatics vol. 7 20. 20 May. 2015, doi:10.1186/s13321-015-0069-3

Appendix

Table 1: **Drugs from the 9 clusters that most related to COVID-19 cluster**

alclometasone dipropionate	Adenosine Diphosphate
1-(2-bromophenyl)-3-(7-cyano-3H-benzotriazol-4-yl)urea	2-Acetylaminofluorene
13-hydroxy-10-oxo-11-octadecenoic acid	4-oxo-2-nonenal
8-chloro-2'-deoxyguanosine	Adenosine Triphosphate
Aldosterone	2-Aminopurine
15-acetyldeoxynivalenol	2-((aminocarbonyl)amino)-5-(4-fluorophenyl)-3-thiophenecarboxamide
8-chlorodeoxyadenosine	4-oxoretinoic acid
alfacalcidol	Aflatoxin B1
8-Hydroxy-2'-Deoxyguanosine	6-Ketoprostaglandin F1 alpha
Alitretinoin	aflatoxin G1
1-Methyl-3-isobutylxanthine	4-(N-methyl-N-nitrosamino)-1-(3-pyridyl)-1-butanone
18alpha-glycyrrhetinic acid	7-ketocholesterol
9-(2-hydroxy-3-nonyl)adenine	AH 23848
alkannin	5-chloro-2'-deoxycytidine
1-Naphthylisothiocyanate	2-morpholin-4-yl-6-thianthren-1-yl-pyran-4-one
22-hydroxycholesterol	8-epi-prostaglandin F2alpha
9-methoxycamptothecin	AICA ribonucleotide
alpha-chaconine	6-((2-(4-imidazolyl)ethyl)amino)heptanoic acid 4-toluidide
1-nitropyrene	4-cresol sulfate
24-hydroxycholesterol	8-hydroxyeicosatetraenoic acid
abacavir	Amlodipine
alpha-damascone	4-phenylbutyric acid
1-trifluoromethoxyphenyl-3-(1-propionylpiperidine-4-yl)urea	4'-methoxychalcone
24-norursodeoxycholic acid	9-deoxy-delta-9-prostaglandin D2
Acetyl Coenzyme A	Amobarbital
Alprostadil	4-tolyl isocyanate
2-(2-aminoethyl)pyridine	4-methylbenzaldehyde
25-hydroxycholesterol	abamectin
acyline	amorolfine
Amphotericin B	4-bromophenacyl bromide

Drug graph with threshold 0.5

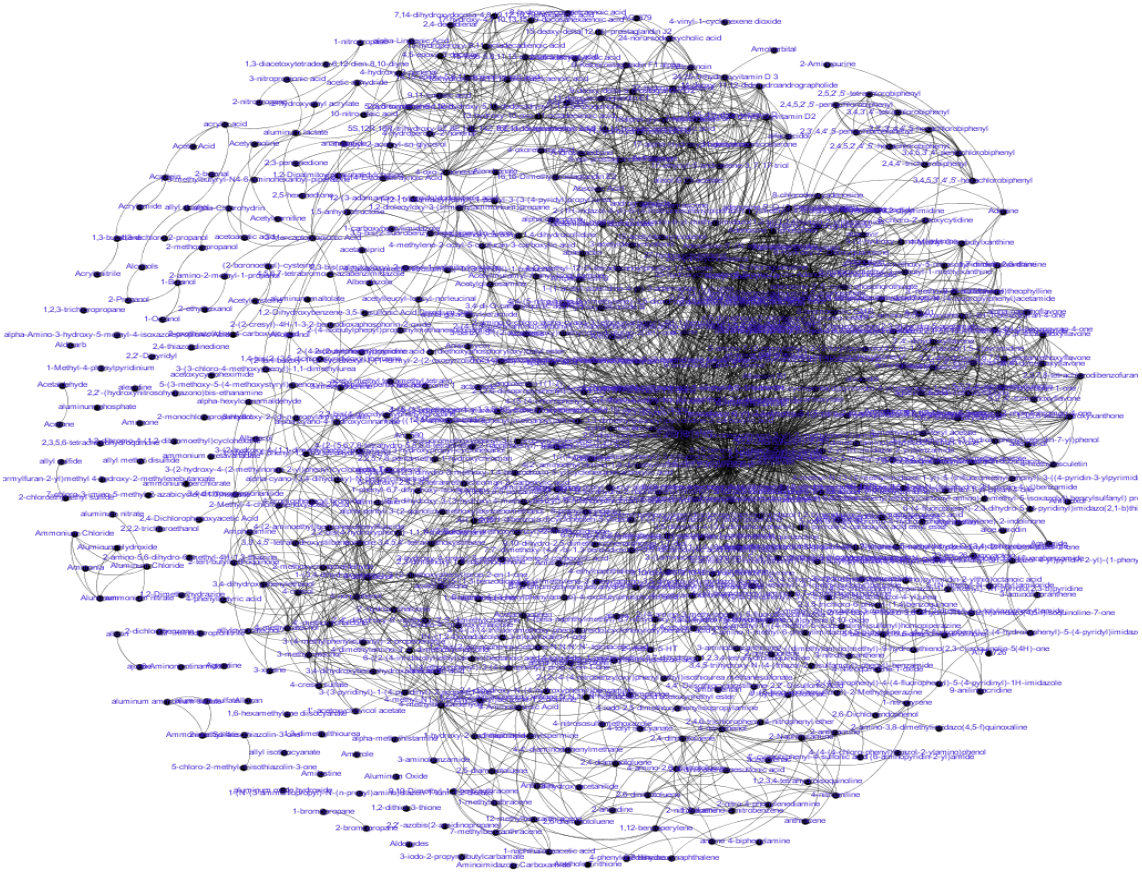

Table 2: **Disease in the COVID-19 cluster**

Lupus Erythematosus	Neuralgia	Coma	Mouth Neoplasms
Amyotrophic Lateral Sclerosis	Neurogenic Inflammation	Calcinosis	Pleural Effusion
Glomerulonephritis	Berylliosis	Respiratory Hypersensitivity	Hypokalemia
Rhabdomyolysis	Polyuria	Neuromuscular Manifestations	Pulmonary Emphysema
Hyperalgesia	Leukopenia	Death	Eye Diseases
Eosinophilia	Multiple Organ Failure	Postoperative Complications	Kidney Tubular Necrosis
Cardiovirus Infections	Acute Lung Injury	Coronary Restenosis	Cachexia
Meningitis	Hyperkalemia	Vascular Diseases	Multiple Myeloma
Cerebrovascular Disorders	Stevens-Johnson Syndrome	Schizophrenia	Hernia
Anaphylaxis	Hearing Loss	Exanthema	Angioedema
Peptic Ulcer	Stomach Ulcer	Pleural Diseases	Behcet Syndrome
Diabetic Angiopathies	Neutropenia	Fibromatosis	Hyperoxaluria
Glomerulonephritis	Hypoglycemia	Purpura	Emphysema
Soft Tissue Neoplasms	Respiratory Syncytial Virus Infections	Bronchopulmonary Dysplasia	Hemolysis
Multiple Sclerosis	Fanconi Anemia	Nephritis	Hepatolenticular Degeneration
Severe Acute Respiratory Syndrome	Pancytopenia	Coxsackievirus Infections	Catalepsy
Respiratory Tract Diseases	HIV Infections	Cardiotoxicity	Carotid Artery Diseases
Chlamydia Infections	Cardiomyopathies	Brain Injuries	Diabetic Retinopathy
Leishmaniasis	Fever	Placenta Diseases	Pruritus
Aneurysm	Myocarditis	Polymyositis	Autoimmune Diseases
Glomerulosclerosis	Cardiovascular Diseases	Dermatitis	Brain Ischemia
Purpura	Anemia	Cystic Fibrosis	Spinal Cord Compression
Cholestasis	Thyroid Neoplasms	Azotemia	Silicosis
Tobacco Use Disorder	Pulmonary Fibrosis	Airway Obstruction	Nephritis
Abortion	COVID-19	Torsades de Pointes	Cystitis
Hearing Disorders	Tachycardia	Hepatitis C	Hepatic Encephalopathy
Cholangiocarcinoma	Venous Thromboembolism	Asthma	Leishmaniasis
Fatigue	Acquired Hyperostosis Syndrome	Nephrotic Syndrome	Pulmonary Edema
Liver Failure	Osteolysis	Mesothelioma	Vasculitis
Ureteral Obstruction	Pulmonary Disease	Hypothermia	

Drug graph with threshold 0.7