Clinical Phenotypes in Long COVID: Unsupervised Clustering of Multi-Omics Profiles

Alan Wu, Annie Cao, Carol Zhou

Introduction

Coronavirus disease (COVID-19), caused by the SARS-CoV-2 virus, is the deadliest outbreak of the 21st century. Although COVID-19 vaccines have significantly reduced the risk of severe disease and mortality, research indicates that 10%-30% of individuals continue to experience persistent symptoms following their acute phase of infection, typically after 3 months [1]. This condition is known as the post-acute sequelae of COVID-19 (PASC), commonly referred to as “long COVID,” which is a multisystemic condition characterized by a range of symptoms that persist long after the acute phase of SARS-CoV-2 infection. These symptoms include myalgia, headache, palpitations, tinnitus, and dyspnea, which collectively have profound impacts on patients’ quality of life [2]. The persistence of these long-term effects has been described as a potential public health crisis [3]. Since long COVID is a newly emerging set of conditions, understanding the physiologic mechanisms holds the potential to transform clinical healthcare strategies, possibly improving diagnostics and treatments for individuals affected by this syndrome.

Our project builds upon the foundational work of Wang et al., titled “Sequential multi-omics analysis identifies clinical phenotypes and predictive biomarkers for long COVID [1].”
In their study, plasma samples from 117 hospitalized COVID-19 patients were analyzed during the acute infection and 6-month post-infection stage. They measured a comprehensive set of 782 biomarkers, including 47 cytokines, 274 proteins, and 635 metabolites. They then used an autoencoder and k-means clustering to identify the changes in these biomarkers between the two phases. Three distinct disease phenotypes were found through unsupervised clustering of multi-omics profiles, each associated with varying symptom severity. Specifically, Cluster A demonstrated minimal molecular deviations and symptoms, representing a relatively mild phenotype. Cluster B was characterized by predominant triglycerides and organic acid structures, indicating distinct metabolic disruptions. Cluster C, the most heterogeneous group, exhibited pronounced inflammatory and metabolic markers, reflecting a more severe and complex clinical profile. While this work offered valuable insights, it had limitations in the selection of clustering algorithms and validation metrics, leaving space for methodological improvement. Notably, 1) the tested range of clusters (k) was unspecified, 2) silhouette score was the only clustering metric utilized which presents possible biases toward certain cluster types, 3) no silhouette score values were provided to evaluate the efficacy of clustering, 4) early stopping was not effectively implemented, and 5) no validation set was used which could lead to overfitting and the inability to choose the dimensionality of the latent representation.

In this project, we aim to investigate distinct clinical phenotypes of long COVID-19. Using the multi-omics profiles, including cytokine, proteomic, and metabolomic data, collected during the acute infection and the 6-month post-infection stage, we employ a more rigorous unsupervised clustering pipeline. In addition to replicating their results, we implemented a variational autoencoder (VAE) and Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction, as well as spectral and agglomerative clustering for phenotyping. The clustering results were further evaluated by Silhouette, Calinski-Harabasz, and Davies-Bouldin scores, ensuring a robust clustering evaluation. Clustering metrics were scaled and aggregated in order to select the most optimal clustering. Overall, our methodology improves the rigor of the analysis, providing deeper insights into the heterogeneity and long-term impacts of post-acute COVID-19 symptoms.

Methods

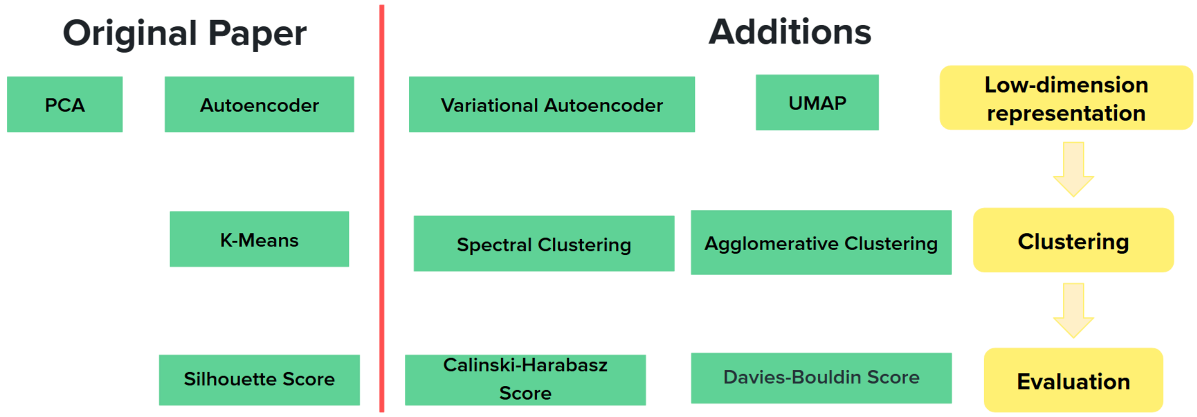


Figure 1. Method overview of dimensionality reduction, clustering, and evaluation pipeline.

Data Preparation

We download the data from the supplemental of the article's protocol [4]. Following their Part 4 instructions, we imported the data file, split it into different datasets based on the type of molecule (cytokine, protein, and metabolite) and the sampling time (acute and long COVID), and calculated the difference between acute and convalescence values, which was used for all later analyses.

Low-dimensional Representation

To replicate the original paper's result, we first utilized Principal Component Analysis (PCA) as a dimensionality reduction technique. PCA linearly transforms the data onto a new coordinate system such that the directions capturing the largest variances in the data can be identified [5]. To assess the efficiency of the dimensionality reduction, we plotted the normalized cumulative sum of variance explained by each principal component.

We then reproduced the paper's autoencoder for a non-linear dimensionality reduction approach. An autoencoder is a type of neural network designed to learn efficient representations of data by encoding it into a lower-dimensional space and then reconstructing the original input from this compressed representation. It aims to minimize the difference between the inputted data and the outputted reconstruction [6]. According to the paper's protocol, the autoencoder reduces the data from 782 features/dimensions to 30 features. For the encoder, the number of neurons in each hidden layer is 100, 70, 50, and 30. For the decoder, the number of neurons in each hidden layer is 50, 70, and 100. The input and output layers both have a dimension of 782 (Figure 2). The model was trained for 1000 epochs, with an early stopping implemented if the loss does not improve for 1000 consecutive epochs.

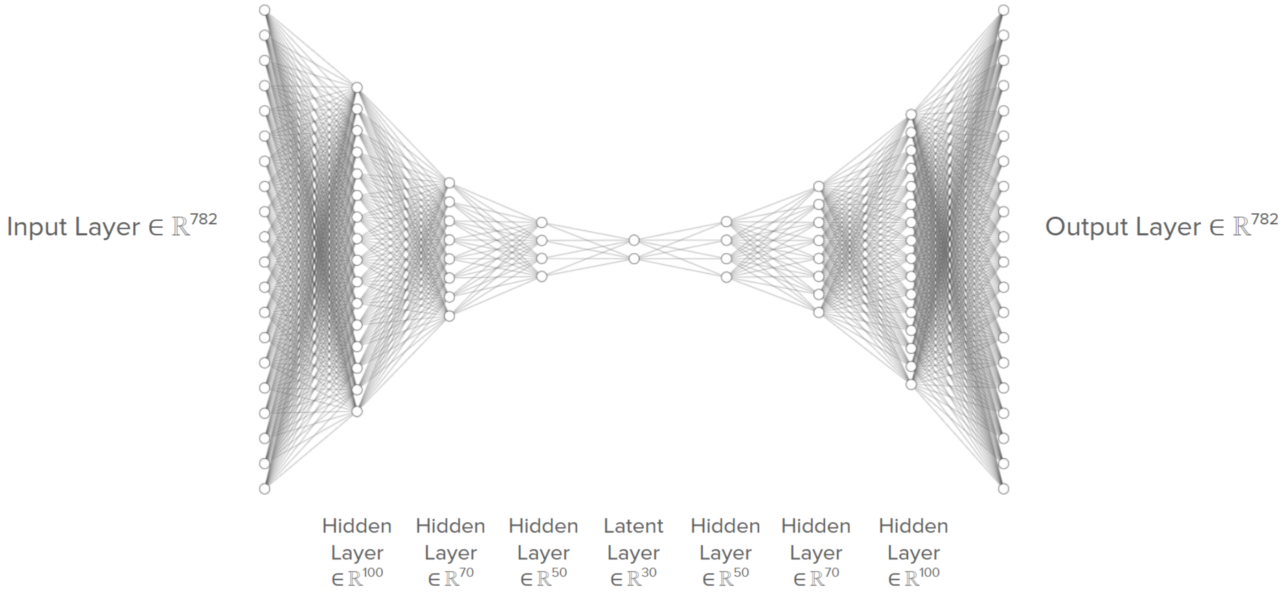

Input Layer ∈ $\mathbb{R}^{782}$
Hidden Layer ∈ $\mathbb{R}^{100}$
Hidden Layer ∈ $\mathbb{R}^{70}$
Hidden Layer ∈ $\mathbb{R}^{50}$
Latent Layer ∈ $\mathbb{R}^{30}$
Hidden Layer ∈ $\mathbb{R}^{50}$
Hidden Layer ∈ $\mathbb{R}^{70}$
Hidden Layer ∈ $\mathbb{R}^{100}$
Output Layer ∈ $\mathbb{R}^{782}$

Figure 2. Original paper's autoencoder architecture.

For our autoencoder, we implemented VAE, a neural network that could compress the data into a probabilistic latent space and reconstruct it. This enables the capture of complex and non-linear patterns. VAE has shown promise in reducing dimensionality while maintaining low reconstruction loss more effectively than traditional autoencoders. We attempted to utilize these strengths to create a low-dimension latent representation that is more representative and lower dimension than the paper's autoencoder.

5-fold cross-validation sets were created in order to prevent overfitting and allow for the determination of the dimensionality of the latent layer. We trained our VAE on the 5-fold cross-validation for n $\in$ {2, 4, 6, 8, 10, 12, 14, 16} dimensional latent layers. For the encoder, the number of neurons in each layer is 512, 128, 32, and 2n (for the mean and variance layer). For the decoder, the number of neurons in each layer is 32, 128, 512. The input and output layers were both 782 neurons (Figure 3). The SeLU activation function was used for all fully connected layers except the output layer in which the Sigmoid activation function was used. The model was trained for a maximum of 100 epochs, with an early stopping implemented if the loss does not improve for 10 consecutive epochs.

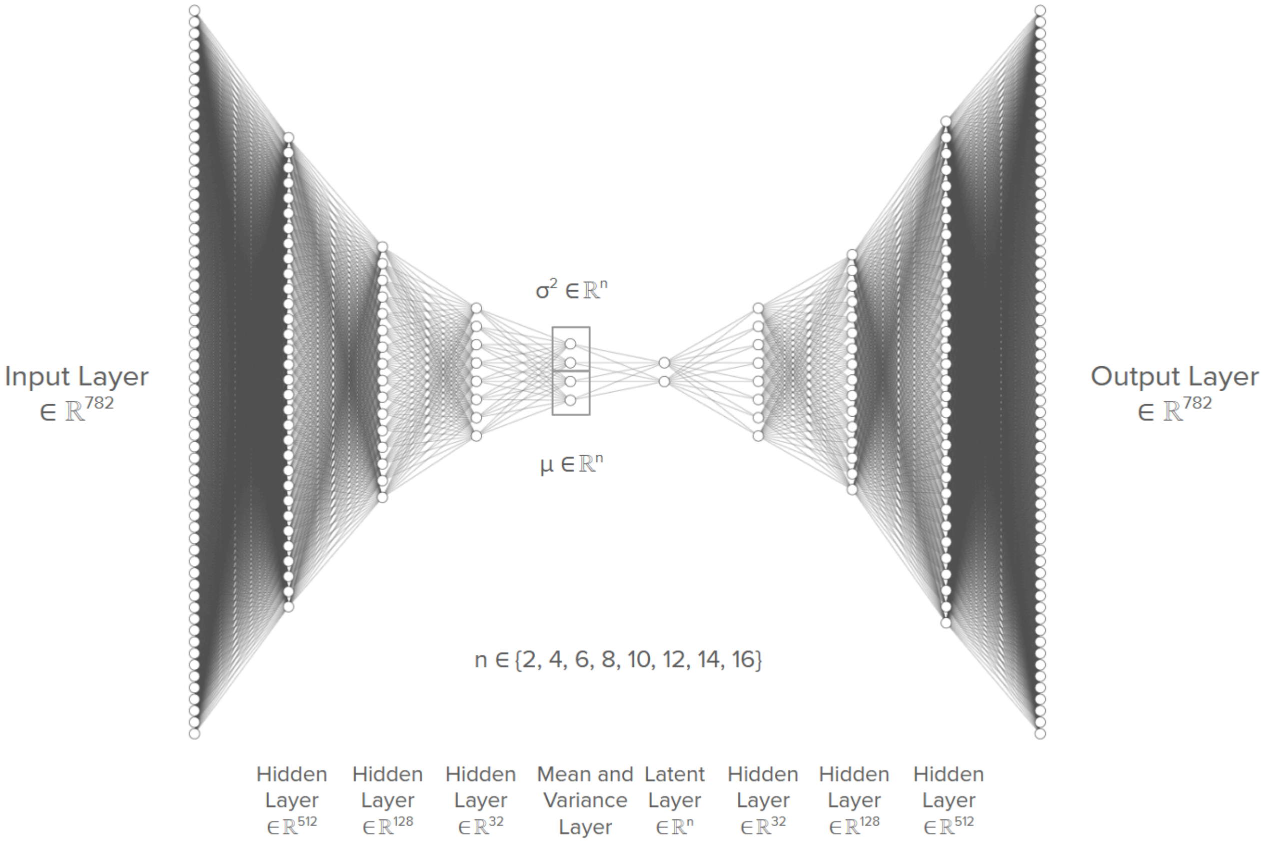

Input Layer $\in \mathbb{R}^{782}$
$\sigma^2 \in \mathbb{R}^n$
$\mu \in \mathbb{R}^n$
Output Layer $\in \mathbb{R}^{782}$
$n \in \{2, 4, 6, 8, 10, 12, 14, 16\}$
Hidden Layer $\in \mathbb{R}^{512}$
Hidden Layer $\in \mathbb{R}^{128}$
Hidden Layer $\in \mathbb{R}^{32}$
Mean and Variance Layer
Latent Layer $\in \mathbb{R}^n$
Hidden Layer $\in \mathbb{R}^{32}$
Hidden Layer $\in \mathbb{R}^{128}$
Hidden Layer $\in \mathbb{R}^{512}$

Figure 3. Variational Autoencoder Architecture.

We further applied UMAP, which is a non-linear dimensionality reduction technique that attempts to find a low dimensional projection that has the closest fuzzy topological structure as the manifold in the high dimensional space, in order to reduce the dimensions to 2 [7]. We decided on 2 dimensions due to the practicality of visualization and the rarity that a dimension greater than 2 is necessary for UMAP with real-world data structures.

Clustering

The initial analysis proceeded with K-means clustering, which partitions data points into a predefined number of clusters by minimizing the within-cluster variance (squared distance), based on the data similarity [8].

The general algorithm for K-means is as follows:

1. K centroids are initiated
2. Points are assigned to a centroid based on a distance metric
3. Centroids are recalculated based on the assigned points
4. Repeat steps 2-3

We further utilized spectral clustering, which constructs graph representations of the data (e.g., using a nearest-neighbor graph) and identifies clusters based on the eigenvalues of the graph's Laplacian matrix. We used the radial basis function as the affinity. This method works effectively for scenarios with a few clusters of relatively even sizes and non-flat geometrics [9].

Spectral clustering has three main steps:

1. Build a similarity graph
2. Project the data into lower dimensional space
3. Cluster the data using a traditional clustering algorithm such as K-means

Finally, we applied agglomerative clustering, a bottom-up hierarchical approach in which each observation starts at its own cluster, and clusters are successively merged based on a linkage criterion [10].
The general algorithm for agglomerative clustering is as follows:

1. Initialize each data point as its own cluster
2. Merge the two closest clusters based on distance
3. Update distances between clusters
4. Repeat steps 2-3

For all the clustering methods, we evaluated the number of clusters (*k*) ranging from 2 to 6. For agglomerative clustering, we simply extracted the clusters from the hierarchical structure. This range was selected to balance the need for meaningful groupings with the constraints of the limited number of patients. A cluster with only a few patients is not meaningful. In addition to not attempting to cluster above 6 clusters, any clustering method and representation combination that resulted in a cluster with less than 10 patients was discarded.

<u>*Cluster Evaluation Metrics*</u>

To analyze the clustering results, we used the Silhouette, Calinski-Harabasz, and Davies-Bouldin scores.

The Silhouette Score measures the intra-cluster coherence and inter-cluster separation. It is calculated by the mean value of $s(i) = \frac{b(i)-a(i)}{max(a(i), b(i))}$, for each point i in a cluster. $b(i)$ is the average distance from point i to all points in the nearest cluster that is not its own. $a(i)$ is the average distance from point i to all other points in its own cluster [11]. The Silhouette Score ranges between -1 and 1. A value near +1 means the data point is in the correct cluster, near 0 means the data point might belong in some other cluster, and near -1 means the data point is in the wrong cluster.

The Calinski-Harabasz Score, also known as the Variance Ratio Criterion, measures how well clusters are separated from each other in a dataset. It is calculated:

$$CH = \frac{N-K}{K-1} \cdot \frac{tr(B_K)}{tr(W_K)}$$

$$tr(B_K) = \sum_{a=1}^{K} n_a(C_a - C)(C_a - C)^T$$

$$tr(W_K) = \sum_{a=1}^{K} \sum_{x \in C_a} (X - C_a)(X - C_a)^T$$

N is the number of data points, and K is the number of clusters. n_a is the number of points in cluster A, C_a is the set of points in cluster K and C is the center of the whole dataset. B_k and W_k are the between and within cluster scatter matrices. Between-cluster dispersion is calculated by looking at the distances between the cluster centroids and the global mean of all points. Within-cluster dispersion looks at how much the data points in each cluster deviate from their respective cluster centroids. At last, the higher the Calinski-Harabasz score, the more well-separated and compact the clusters are [12].

The Davies-Bouldin Score measures the average similarity between each cluster and its most similar cluster, where similarity is defined as the ratio of intra-cluster distance to inter-cluster distance. Given a dataset X and k clusters, the Davies-Bouldin Score is calculated:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}(\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i, X_j)})$$

$\Delta(X_k) = \frac{1}{|X_k|}\sum_{x\in X_k} ||x - C_k||$ where $|X_k|$ is the number of points in cluster X_k, and C_k is the centroid of X_k

$\delta(X_i, X_j) = ||C_i - C_j||$ where C_i and C_j are centroids

$\Delta(X_k)$ computes the average distance between points within the cluster and the centroid of the cluster, while $\delta(X_i, X_j)$ measuring the distance between the centroids of the two clusters. Lastly, the Davies-Bouldin score ranges 0 to infinity. The lower the score, the better cluster separation and compactness [13].

We calculated the Silhouette, Calinski-Harabasz, and Davies-Bouldin scores for clustering results obtained using k-means, spectral, and agglomerative clustering, applied to the low-dimensional representations produced by the paper's autoencoder, our VAE, and UMAP. By using multiple evaluation metrics, we aimed to ensure the clustering solution is genuinely robust and provides a comprehensive assessment of the clustering algorithm's performance across multiple aspects of cluster quality.

We then normalized the scores based on the maximum score obtained for each metric. To determine the optimal combination of the number of clusters, clustering method, and low-dimensional representation, we aggregated the scores by adding Silhouette and Calinski-Harabasz (since higher values indicate better clustering) but subtracting Davies-Bouldin (since lower values indicate better clustering).

Biomarker Visualization

In order to determine biomarkers with significant differences, we conducted one-way Analysis of Variance (ANOVA) tests on each biomarker clustering. In order to adjust for conducting 782 tests, we applied a Bonferroni correction.

Thus, $\alpha = 0.05 / 782 = 6.4 \times 10^{-5}$.

95% confidence intervals were also calculated for each biomarker clustering.

Results

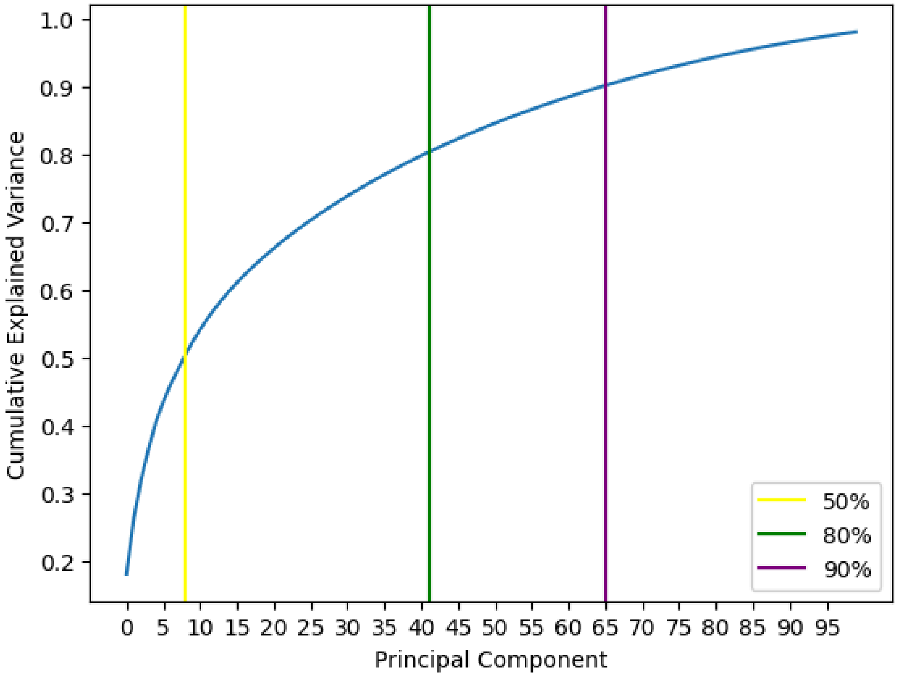


Figure 4. Cumulative explained variance by the number of principal components.

To capture 90% of the variance, 66 principal components are required. To capture 80% of the variance, 42 principal components are required, and to capture 50% of the variance, 9 principal components are required.

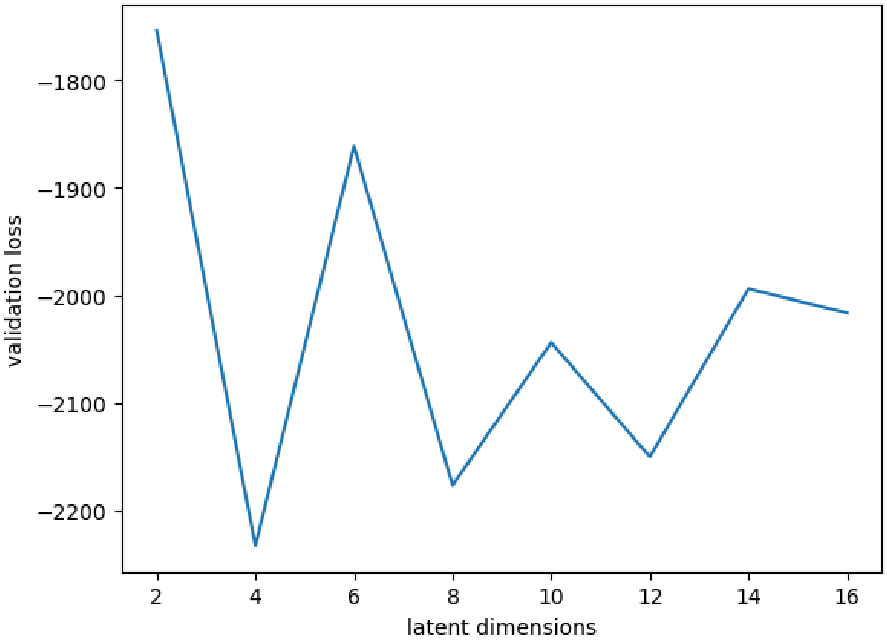


Figure 5. Validation loss across latent dimensions of VAE.

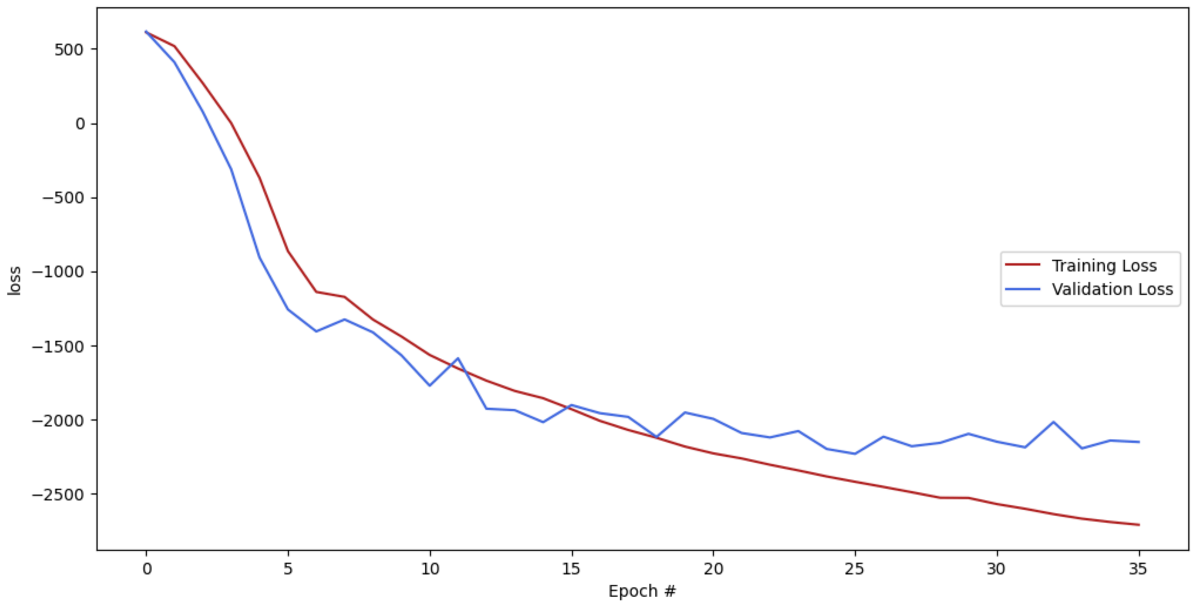


Figure 6. Training and validation loss across epochs of VAE.

The validation loss reached its minimum at 4 latent dimensions (Figure 5), indicating that 4 latent dimensions are optimal for minimizing validation loss. At 2 latent dimensions, the validation loss was significantly higher, suggesting that the model was unable to capture sufficient information in such a reduced latent space. Beyond 4 dimensions, the validation loss exhibited irregular fluctuations, indicating that additional latent dimensions did not consistently improve performance.

Clustering Method	Dimensionality Reduction Technique	Number of Clusters	SS	CH	DB	Scaled SS	Scaled CH	Scaled DB	Scaled Total
k-means	original	4	0.64	386.65	0.42	0.96	1.00	0.47	1.48
agglomerative	original	4	0.62	355.01	0.41	0.93	0.92	0.46	1.38
k-means	original	3	0.67	316.41	0.45	1.00	0.82	0.50	1.32
agglomerative	original	3	0.66	300.04	0.44	0.99	0.78	0.50	1.27
spectral	original	3	0.65	286.52	0.44	0.98	0.74	0.50	1.22
spectral	original	2	0.62	143.08	0.50	0.92	0.37	0.56	0.73
k-means	original	2	0.61	144.64	0.51	0.92	0.37	0.57	0.72
agglomerative	original	2	0.61	144.64	0.51	0.92	0.37	0.57	0.72
spectral	umap	5	0.46	165.65	0.69	0.69	0.43	0.77	0.34
k-means	umap	5	0.45	167.38	0.71	0.68	0.43	0.80	0.32
k-means	umap	4	0.46	164.89	0.74	0.69	0.43	0.83	0.29
spectral	umap	4	0.46	161.77	0.75	0.69	0.42	0.84	0.26
agglomerative	umap	5	0.44	150.32	0.71	0.66	0.39	0.80	0.25
spectral	umap	2	0.45	135.35	0.77	0.68	0.35	0.86	0.17
k-means	umap	2	0.45	145.98	0.82	0.68	0.38	0.92	0.14

agglomerative	umap	4	0.43	150.72	0.81	0.65	0.39	0.91	0.13
k-means	umap	3	0.43	150.17	0.82	0.65	0.39	0.92	0.12
agglomerative	umap	2	0.42	101.67	0.72	0.62	0.26	0.81	0.07
spectral	umap	3	0.40	133.97	0.83	0.60	0.35	0.93	0.01
agglomerative	umap	3	0.38	127.16	0.89	0.57	0.33	1.00	-0.10

Table 1. Comparison of clustering methods, dimensionality reduction techniques, and the number of clusters, with scaled Silhouette Score (SS), Calinski-Harabasz Score (CH), and Davies-Bouldin Score (DB) based on their respective maxima, ranked by scaled total score. Note that some combinations of clustering methods, dimensionality reduction, and number of clusters do not appear because they generated clusters of less than 10 patients. For example, no variational autoencoder method was able to meet this threshold.

The results in Table 1 shows that the combination of k-means clustering, original autoencoder low-dimensional representation, and 4 clusters yields the optimal clustering quality. This configuration achieves the highest total scaled score of 1.48, with the best CH and DB scores. This clustering will be used for the remainder of the paper.

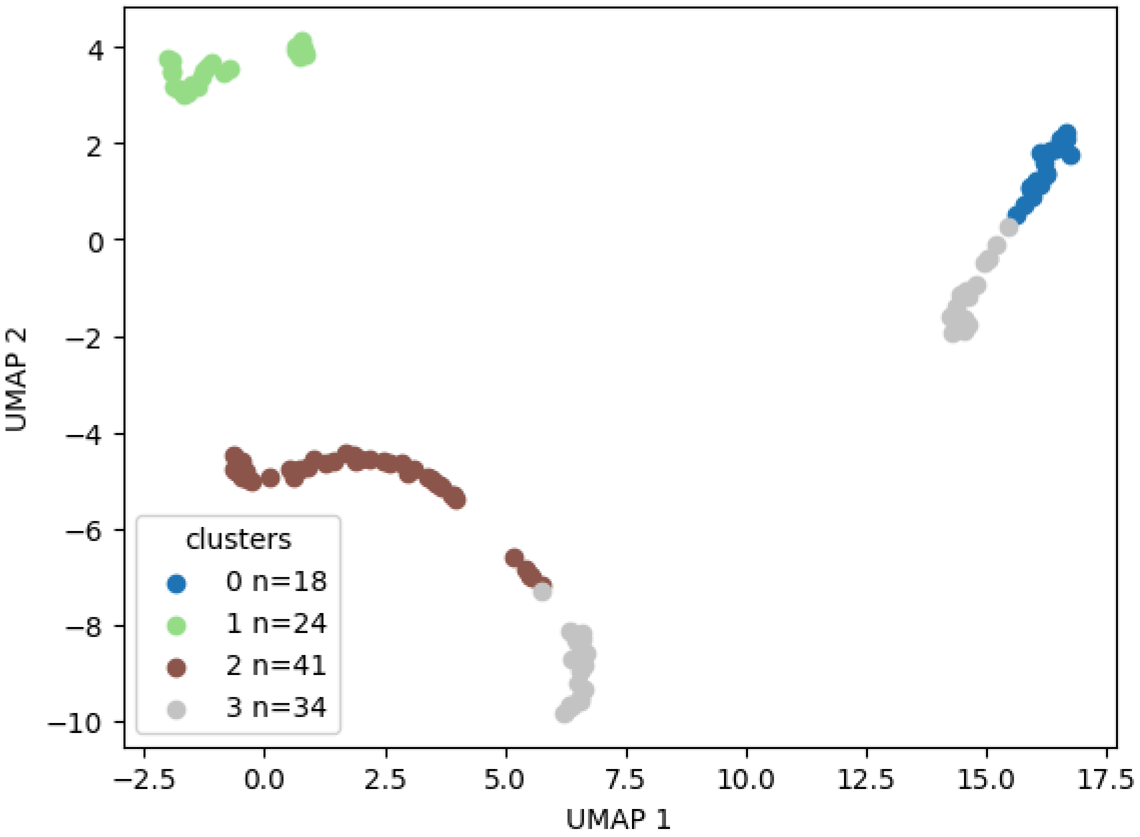
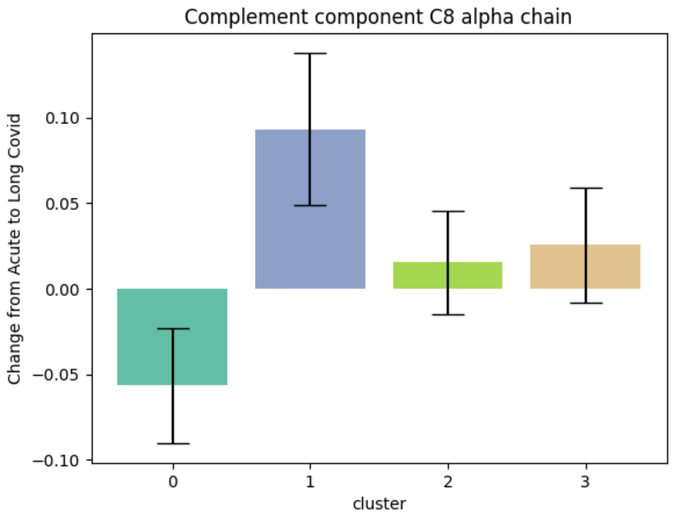
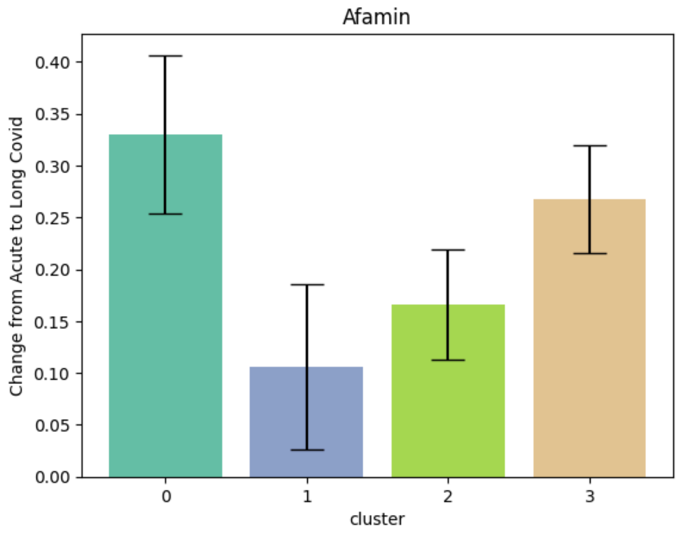
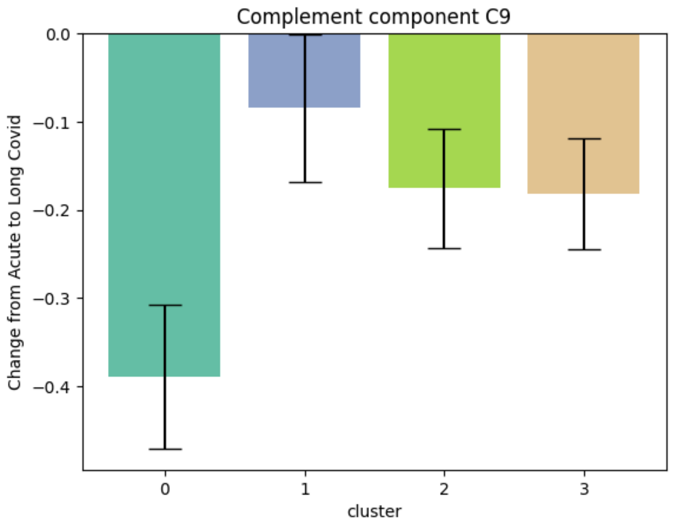


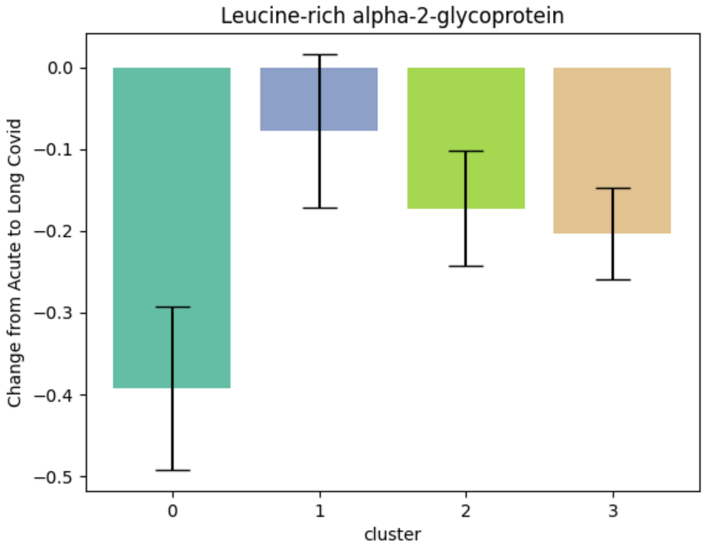
Figure 7. UMAP visualization of clusters (n = 117) in reduced dimensional space.

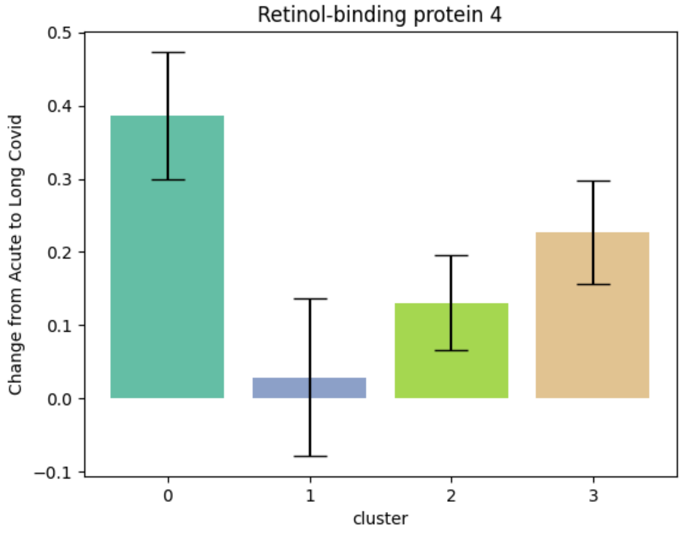
UMAP is used to visualize the 2-dimension representation of the 30-dimensional latent space generated by the original autoencoder. Four clusters were identified, with the size as follows: Cluster 0 (n = 18), Cluster 1 (n = 24), Cluster 2 (n = 41), and Cluster 3 (n = 34).

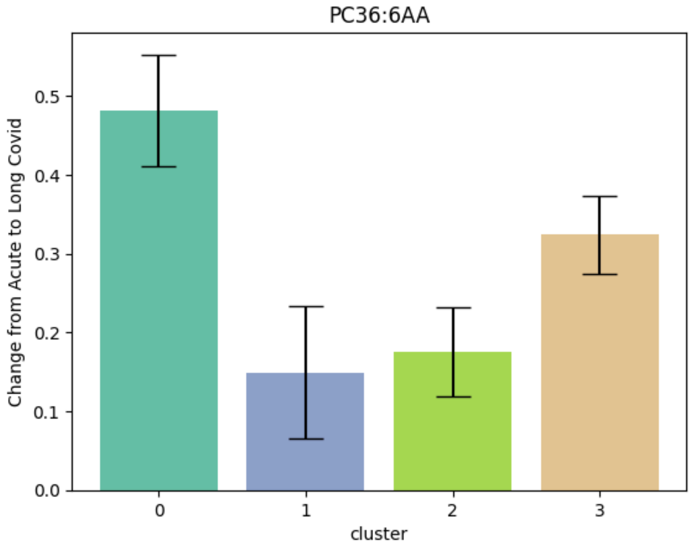

Complement component C8 alpha chain
Change from Acute to Long Covid
0.10
0.05
0.00
−0.05
−0.10
0
1
2
3
cluster


Afamin
Change from Acute to Long Covid
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
cluster


Complement component C9
Change from Acute to Long Covid
0.0
−0.1
−0.2
−0.3
−0.4
0
1
2
3
cluster


Leucine-rich alpha-2-glycoprotein
Change from Acute to Long Covid
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
0
1
2
3
cluster


Retinol-binding protein 4
Change from Acute to Long Covid
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
3
cluster


PC36:6AA
Change from Acute to Long Covid
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
cluster

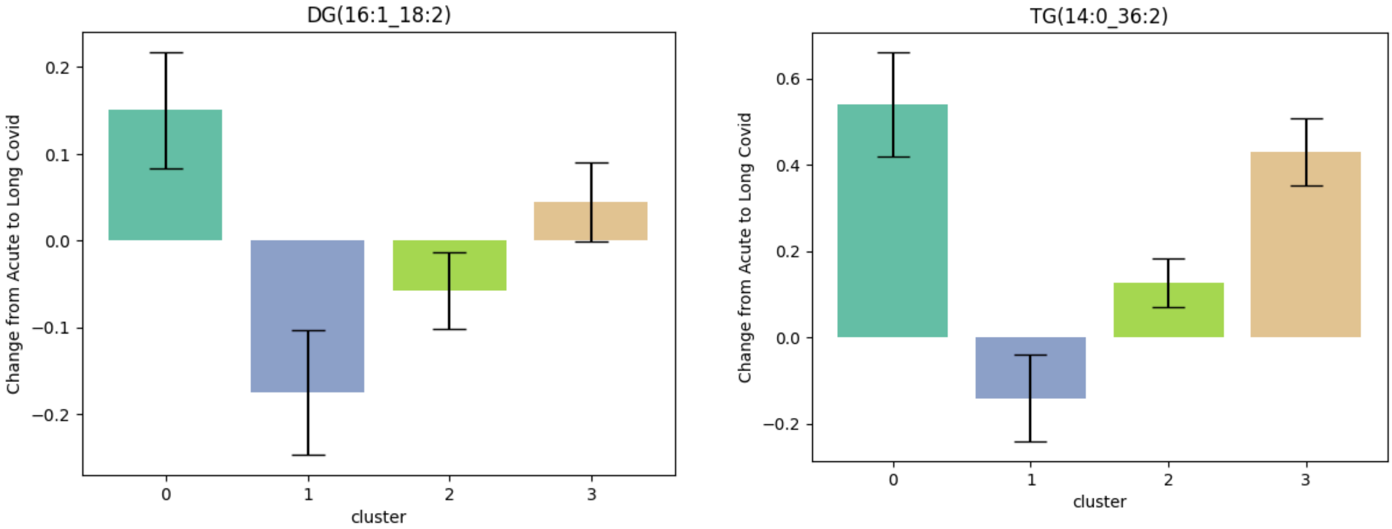


Figure 8. Cluster-specific biomarker changes from acute to long Covid. 235 out of 782 biomarkers had statistically significant clusterings ($p<6.4 \times 10^{-5}$) by one-way ANOVA. 8 biomarkers were manually selected to be representative of the differences between clusters.

PC36:6AA is a specific type of phosphatidylcholine molecule. DG(16:1_18:2) is a specific type of diacylglycerol molecule. TG(14:0_36:2) is a specific type of triglyceride molecule.

Discussion

PCA Verification

To recreate the original paper's results, we first used PCA as a dimensionality reduction approach. According to Figure 4, 66 principal components are required to capture 90% variance, which is impractical given the limited sample size of 117 patients. The finding aligns with the paper's assessment that PCA does not perform well for this dataset, as it struggles to capture the complexity of the multi-omics data in low dimensions. However, PCA demonstrates some utility for preliminary dimensionality reduction, as it could capture 50% of the variance using just 9 components. Despite this, the inability to reduce the dimensions to a more practical level without the significant loss of information underscores the need for more advanced methods.

Original Paper Replication

As shown in Table 1, for cluster numbers running from 3 to 5, the combination of the original autoencoder, k-means clustering, and 3 clusters yields the highest Silhouette score of 0.67, which aligns with the paper's result. However, since the paper did not specify the exact Silhouette score they obtained, though we were able to verify the number of clusters they used, we could not directly assess the quality of their clustering. Moreover, a Silhouette score of 0.67 suggests that the cluster quality is moderate on a scale from -1 to 1.

Comparison to Our Methodology

A limitation we noticed in the paper's methodology is that k-means clustering tends to form spherical clusters with roughly the same size, while in reality, we are unsure whether that reflects the true structure of the data. In addition, the Silhouette score, typically used to evaluate k-means clustering, has a bias toward favoring spherical and regular-shaped clusters. Thus, to address these concerns and ensure the robustness of our results, we employed spectral and agglomerative clustering, which are capable of forming clusters with arbitrary shapes. We also checked cluster quality with Calinski-Harabasz and Davies-Bouldin scores, which could provide further insights about the distinctness and separation of our clusters.

Indicated in Table 1, the combination of k-means clustering, original autoencoder low-dimensional representation, and 4 clusters yields the optimal clustering quality, with the highest total normalized score of 1.48. This configuration also achieves the best Calinski-Harabasz and Davies-Bouldin scores and the 4th best Silhouette score. In contrast, the paper's combination of k-means clustering, original autoencoder, and 3 clusters did not perform as well as the previous combination on the other evaluation metrics. This could indicate that our clustering approach may provide a more nuanced and better-separated representation of the data and that the true structure of the data is more complex than what the paper suggests.

UMAP

Based on the 2-dimensional representation, Cluster 1 demonstrated clear separation from the other clusters in the reduced dimensional space (Figure 7). In contrast, Clusters 2 and 3 exhibited close proximity. Similarly, Cluster 0 also showed proximity to Cluster 3, indicating some level of similarity in their profiles.

VAE

We attempted to use a VAE to improve upon the original paper's autoencoder and create a better low-dimension representation. Besides changing the autoencoder's architecture, we also made other improvements. We set early stopping with a patience of 10 so that the network would stop when it failed to improve and not at a set number of epochs. We utilized 5-fold cross-validation in order to prevent overfitting and in order to determine the dimensionality of our low-dimension representation. However, we were not successful in improving clustering results. All clustering combinations with the VAE did not meet the preset criteria that every cluster needed at least 10 patients (Table 1). We hypothesize that the reason why clustering is poor, is due to the inherent Gaussian distribution of the latent space. VAEs encourage the formation of a Gaussian latent space through the Kullback-Leibler divergence loss. While this has been shown to be useful in terms of reconstruction with a relatively low-dimension bottleneck as compared to traditional hourglass autoencoders, the clustering structure of the low-dimension representation may be poor as a compromise. From our limited testing, the Gaussian nature of the low-dimension representation encourages the formation of one large cluster with the rest of the clusters having very few data points.

Clinical Analysis

Cluster 0 showed decreases in complement component C8 alpha chain, complement component C9, and leucine-rich alpha-2-glycoprotein. This suggests immune dysfunction from possible chronic inflammation due to PASC [14][15]. Elevated afamin, retinol binding protein 4, diglycerol (DG(16:1_18:2)), and triacylglycerol (TG(14:0_36:2)) support this interpretation of systemic inflammation [16][17][18][19]. Elevated phosphatidylcholine and triacylglycerol suggest a lipid metabolic shift possibly from an inflammatory stress response. Due to these factors, cluster 0 is the inflammation cluster.

Cluster 1 showed mild increases in complement component C8 alpha chain which may indicate an immune response [14]. The slight increase in afamin and phosphatidylcholine may indicate early stages of inflammation [16][20]. However, decreases in diglycerol and triacylglycerol indicate compensatory metabolic adjustment [18][19]. Due to these factors, cluster 1 is the mild immune response with metabolic compensation cluster.

Cluster 2 showed mild increases in afamin, retinol binding protein 4, phosphatidylcholine, and triacylglycerol which could indicate metabolic stress [16][17][20][19]. The cluster also showed mild decreases in leucine-rich alpha-2-glycoprotein and complement component C9 which suggest immune downregulation [15][14]. Due to these factors, cluster 2 is the mild immune downregulation and metabolic shift cluster.

Cluster 3 showed similar changes to cluster 0 but less extreme. Therefore, it is the mild inflammation cluster.

Limitations and Future Work

While our study addressed several methodological gaps in the original paper, certain limitations still remain. Our project relied on the data collected by Wang et. al., which was limited to only 117 hospitalized COVID-19 patients. This small sample size restricts the generalizability of our findings. Larger cohorts are needed to validate the observed clusters and assess their clinical relevance across different populations. Furthermore, our analysis was restricted to plasma biomarker data. Future research work could integrate the plasma biomarker data with complementary datasets, including imaging modalities and other detailed clinical records, to provide a holistic understanding of long COVID pathophysiology.

Another limitation of our study was the lack of temporal sampling to capture the dynamic progression of long COVID. Monitoring biomarker evolution across different stages of the condition (e.g., 3–5, 6–8, and 9–12 months post-infection) could reveal predictive markers of symptom persistence. Incorporating time-series clustering methodologies could offer insights into the dynamic progression of long COVID, allowing for the identification of distinct temporal phenotypes. This approach would also enable comparisons between the evolution of different long COVID phenotypes and other chronic post viral conditions.

Different viral strains of the SARS-CoV-2 virus (e.g., Delta, Omicron) and host factors, such as age, sex, and comorbidities, likely influence the onset and progression of long COVID symptoms. Future studies could explore how these variations collectively affect biomarker patterns and phenotypic clustering. Stratifying cohorts by viral strain or patient characteristics could uncover important differences in long COVID manifestations and inform personalized treatment strategies.

In our project, we sought to enhance the original paper's autoencoder by implementing a VAE to generate a more robust low-dimensional representation. Alongside the architectural changes to the autoencoder, we made other methodological improvements. However, despite these efforts, our approach did not yield improved clustering results. One key limitation of VAEs lies in their latent space design. By enforcing a Gaussian prior, VAEs often produce overly smooth and continuous latent spaces, which can result in hyperspherical embeddings and poorly defined clusters.

To address these shortcomings, we propose exploring Variational Deep Embedding (VaDE) in future research. VaDE is an advanced unsupervised generative clustering approach that integrates the VAE framework with a Gaussian Mixture Model (GMM) for clustering tasks [21]. By replacing the single Gaussian prior in VAEs with a mixture of Gaussians, VaDE naturally aligns the latent space with clustering objectives. This adjustment enables the generation of distinct and well-separated clusters, making it more suitable for tasks requiring precise phenotypic differentiation [21]. To the best of our knowledge, no prior studies in this field have applied VaDE to identify clinical phenotypes of long COVID when using unsupervised learning algorithms. Future work could pioneer this application, leveraging VaDE's capabilities to identify the nuanced phenotypic clusters and provide deeper insights into the heterogeneity of long COVID.

Implications

Our findings build on existing literature suggesting that metabolic dysregulation plays a central role in the pathophysiology of long COVID. Consistent with prior studies, we observed mean increases in specific metabolites among long COVID patients, indicating sustained disruptions in the metabolic pathways. For instance, one study reported increased activity of the kynurenine pathway in COVID-19 patients, evidenced by elevated kynurenine levels and a higher kynurenine-to-tryptophan ratio, particularly in severe cases [22]. Another study identified elevated levels of inflammatory markers, including phenylalanine, suggesting that metabolic alterations may contribute to the inflammatory processes associated with the disease [23]. Additionally, research has documented significant changes in amino acid metabolism, including increased kynurenine levels, which were linked to immune responses and may influence disease progression [24]. These findings align with our results, reinforcing the idea that metabolic disruptions are central to long COVID's pathophysiology. This convergence highlights the need for further research into metabolic pathways as potential diagnostic and therapeutic targets.

Conclusion

We performed dimensionality reduction using the autoencoder from the original paper, a VAE, and UMAP. Subsequently, we applied k-means, spectral, and agglomerative clustering to the resulting low-dimensional representations. Cluster quality was then evaluated using Silhouette, Calinski-Harabasz, and Davies-Bouldin scores to identify the best combination of models for further analyses. We were able to replicate the original paper's results, but suggested a separate clustering. Clusters were analysed clinically and complex combinations of immune and metabolic changes were found. In order to further explore the complex issue of PASC, we proposed future clinical and technical work.

References

1. Wang, K., Khoramjoo, M., Srinivasan, K., Gordon, P. M. K., Mandal, R., Jackson, D., et al. (2023). Sequential multi-omics analysis identifies clinical phenotypes and predictive biomarkers for long COVID. CR Med 4. doi: 10.1016/j.xcrm.2023.101254
2. Davis, H. E., McCorkell, L., Vogel, J. M., and Topol, E. J. (2023). Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol 21, 133–146. doi: 10.1038/s41579-022-00846-2
3. Ballering, A. V., Zon, S. K. R. van, Hartman, T. C. olde, and Rosmalen, J. G. M. (2022). Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study. The Lancet 400, 452–461. doi: 10.1016/S0140-6736(22)01214-4
4. Khoramjoo, M., Srinivasan, K., Wang, K., Wishart, D., Prasad, V., and Oudit, G. Y. (2024). Protocol to identify biomarkers in patients with post-COVID condition using multi-omics and machine learning analysis of human plasma. STAR Protocols 5, 103041. doi: 10.1016/j.xpro.2024.103041
5. Shlens, J. (2014). A Tutorial on Principal Component Analysis. doi: 10.48550/arXiv.1404.1100
6. Zürn, J. (2019). But what is an Autoencoder? Medium. Available at: https://jannik-zuern.medium.com/but-what-is-an-autoencoder-26ec3386a2af.
7. McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018
8. J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, vol. 5.1, pp. 281–298, Jan. 1967.
9. "SpectralClustering," *scikit-learn*, 2019. https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.SpectralClustering.html
10. "2.3. Clustering," *scikit-learn*. https://scikit-learn.org/1.5/modules/clustering.html
11. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. doi: 10.1016/0377-0427(87)90125-7
12. Łukasik, S., Kowalski, P. A., Charytanowicz, M., and Kulczycki, P. (2016). Clustering using flower pollination algorithm and Calinski-Harabasz index., in 2016 IEEE Congress on Evolutionary Computation (CEC), 2724–2728. doi: 10.1109/CEC.2016.7744132
13. Davies-Bouldin Index (15:08:43+00:00). GeeksforGeeks. Available at: https://www.geeksforgeeks.org/davies-bouldin-index/.
14. Janeway CA Jr, Travers P, Walport M, et al. Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; 2001. The complement system and innate immunity. Available from: https://www.ncbi.nlm.nih.gov/books/NBK27100/
15. Camilli, C., Hoeh, A.E., De Rossi, G. et al. LRG1: an emerging player in disease pathogenesis. J Biomed Sci 29, 6 (2022). https://doi.org/10.1186/s12929-022-00790-6

16. Hans Dieplinger, Benjamin Dieplinger, Afamin — A pleiotropic glycoprotein involved in various disease states, Clinica Chimica Acta, Volume 446, 2015, Pages 105-110, ISSN 0009-8981, https://doi.org/10.1016/j.cca.2015.04.010.
17. Steinhoff Julia S. , Lass Achim , Schupp Michael, Biological Functions of RBP4 and Its Relevance for Human Diseases, Frontiers in Physiology, 12, 2021, https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2021.659977, 10.3389/fphys.2021.659977, 1664-042X
18. van der Veen, J. N., Kennelly, J. P., Wan, S., Vance, J. E., Vance, D. E., & Jacobs, R. L. (2017). The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. Biochimica et biophysica acta. Biomembranes, 1859(9 Pt B), 1558–1572. https://doi.org/10.1016/j.bbamem.2017.04.006
19. Diacylglycerol, when simplicity becomes complex, Carrasco, Silvia et al., Trends in Biochemical Sciences, Volume 32, Issue 1, 27 - 36
20. Staff, M. C. (2022, September 3). *Can triglycerides affect my heart health?* https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/triglycerides/art-20048186
21. Z. jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, “Variational Deep Embedding: A Generative Approach to Clustering.,” *arXiv (Cornell University)*, Nov. 2016.
22. A. F. Almulla, T. Supasitthumrong, C. Tunvirachaisakul, A. A. A. Algon, H. K. Al-Hakeim, and M. Maes, “The tryptophan catabolite or kynurenine pathway in COVID-19 and critical COVID-19: a systematic review and meta-analysis,” *BMC Infectious Diseases*, vol. 22, no. 1, Jul. 2022, doi: https://doi.org/10.1186/s12879-022-07582-1.
23. M. Gietl *et al.*, “Laboratory parameters related to disease severity and physical performance after reconvalescence of acute COVID-19 infection,” *Scientific Reports*, vol. 14, no. 1, May 2024, doi: https://doi.org/10.1038/s41598-024-57448-6.
24. T. Thomas *et al.*, “COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status,” *JCI Insight*, vol. 5, no. 14, Jul. 2020, doi: https://doi.org/10.1172/jci.insight.140327.

Appendix

Code:
https://colab.research.google.com/drive/15K2ohPo8rvPO5UjEW2IzO3LTazSZttFo?usp=sharing