1 Title

Data Imputation and Graph-based Approach for Post-Stroke Immune System Response Profiling

2 Group Members

Jeongho Chae (PID: 730842251)
Dzung Dinh (PID: 730798415)
Sajan Patel (PID: 730466308)
Tripp Whaley (PID: 720512632)

3 Abstract

Strokes (cerebrovascular accidents) can cause a significant, long-term decline in cognitive function (Pendlebury and Rothwell 2009). A recent study utilizing mass cytometry found that the immune system's acute response to stroke is strongly correlated. However, due to the patient's substantial dropout during the study, the relationship between immune system response and cognitive function post-stroke has been underexplored. Consequently, the conclusions drawn from such incomplete datasets might be biased (Enders 2022). To address these limitations, this project focuses on data imputation, which will be further advanced for future works in biomarker discovery.

4 Introduction

4.1 Problem Motivation

Stroke is one of the leading causes of cognitive impairment and dementia; however, the studies of stroke have not yet been well understood. A recent study utilizing mass cytometry found that the immune system's acute response to stroke is strongly correlated with cognitive decline at days 90 and 365 post-stroke (Tsai et al. 2019). However, in such studies, the patient often misses a follow-up visit. Conventional strategies, such as ignoring the missing data or naive imputation (e.g., mean or median), discard useful information from the data, resulting in biased and incomplete conclusions.

To address these limitations, our project targets the data imputation problem. We focus on using EnGen, a state-of-the-art imputation model for high-dimensional mass-cytometry. EnGen is effective at learning the full pattern of the data and is able to generate a more reasonable reconstruction compared to naive imputation. Note that we do not have access to the label (e.g., the Montreal Cognitive Assessment scores) of the dataset, and hence, further analysis on the generated dataset is limited. That being said, we will perform various assessments in this project to confirm that the reconstruction is reasonable. By reconstructing the missing patients, we hope to produce unbiased data for future biomarker discovery.

4.2 Previous work focused on solving this problem

The main study (Tsai et al. 2019) we try to recreate here made novel discoveries around immune system responses at various time points post-stroke (days 2, 5, and 90). In their study, they trained models at each timepoint in conjunction with changes in MoCA scores from days 90 to 365 to model how immune responses at day 2 can provide a clinical indicator as to the extent of cognitive decline a year later. Their approach used an ElasticNet model, which automatically prioritizes the most important features and regularizes less important features, after manually gating for selected cellular features.

Other previous works on studying post-stroke responses have either performed analysis on complete-case analysis or simple imputation, such as mean/median, or interpolation. These methods remain the marginal statistics of the data, but cannot capture the complicated patterns of the data.
As for data imputation, we attempt to utilize an unexplored option through EnGen. EnGen is an Encoder-Decoder deep neural network that utilizes patient samples from a beginning timepoint to generate a similarly sized sample of realistic single cell data at a following time point. In their paper, they examine patients' single-cell samples pre-operation and one hour post-operation.

4.3 Limitations of previous work

The previous post-stroke immune profiling study's limitations were not necessarily in the approach itself but in issues common to many clinical studies - patient dropout. We note that the major days the previous study found as having clinically predictive capabilities had 24, 7, and 14 patients present for blood samples, respectively, and the most notable clinically predictive for cognitive decline being at day 2. We hoped to uncover other potentially predictive biomarkers at timepoints where fewer than 7 patients reported. We note in particular that days 14 and 30 had two and three patient samples available, respectively, and we focus our imputation techniques on these days. Additionally, where the prior study utilized manual gating with an ElasticNet model, we hoped to find predictive biomarkers without any explicit manual gating of cell types necessary.
As for EnGen, their limitations were that the study was more focused on a narrow time-range following a biologically traumatic surgery. We aimed to apply this same methodology to significantly longer timeframes with less patients available for the model to train on.

5 Statement of Contributions

Our key contributions are as follows:

- We use EnGen to reconstruct data for days where the percentage of patient dropout is significantly high (e.g., 3 out of 25 patients reporting).
- We evaluate the quality of generated data to confirm its validity (or lack thereof) for our sample size.
- We perform graph clustering on the *original* data at each timepoint, showing a similar number of communities with correlated immune features in the stroke recovery paper.
- We perform graph clustering on the *imputed* data at relevant timepoints (days 14 and 30), and show that a lack of varied data from patients makes training an effective model from a very limited cohort of patients challenging.

Our code is available at https://github.com/TrippWhaley/comp-bio-final.

6 Methods

6.1 Notation

EnGen. When evaluating EnGen generation quality, we quantify how each single-cell feature's coefficient of variation (ΔCV) changes. The ΔCV of a feature is defined as:

$$CV = \frac{\sigma}{\mu},$$

where σ and μ are the feature's standard deviation and mean, respectively. We first compute ΔCV on the original data. We then concatenate the original with the imputed data (e.g., by EnGen) and compute

$$CV_{concat} = \frac{\sigma_{concat}}{\mu_{concat}}.$$

Finally, we define the change in variability as

$$\Delta CV = |CV_{concat}| - |CV|.$$

Leiden Clustering. For the Leiden Clustering, we model the single-cell data as a graph whose vertices are $V = \{v_1, v_2, ..., v_n\}$, where each node v_j is a 47-dimensional cell profile. Then, we find distinct communities, defined as $C_i \subseteq V$, that optimize the modularity of a partition $P = \{C_1, C_2, ..., C_n\}$.

6.2 Problem Formulation

$$\forall d \in D \qquad P_d := Leiden(V), d \in [1, 2, 3, 5, 7, 14, 30, 90, 365] \tag{1}$$

$$\forall d \in D' \qquad P'_d := Leiden(V'), d \in [14, 30] \tag{2}$$

6.3 Description of your method

To identify immune features through a graph-clustering approach, Leiden clustering was employed on the original CyTOF patient data for all timepoints. Since the CyTOF data included 47 biomarkers, we first performed a Principal Component Analysis (PCA) to reduce the dimensionality of the data down to 20 principal components. Then, since most time points included over 600k cells per patient, we did a geometric sketch (Hie et al. 2019) with a sketch size of 1000 cells to create a representative subset of the patient data while reducing computational demands. With this sketch, we constructed a k-nearest neighbors (kNN) graph and computed a 2D UMAP embedding for later visualization of the graph. Finally, we performed Leiden clustering on the sketched cells to find communities that may correspond to immune features.

To impute features, we train one model for each day using all patients that have data at both Day 1 and Day (14 or 30). We provide analysis on the validity of these imputed features in Section 7 by retraining models on a subset of patients and holding out one patient for a validation set for Day 14 and comparing to a time point with significantly more patients available (Day 2). When including imputed feature data for Days 14 and 30 post-stroke, we employed the exact same computational methods to once again perform Leiden Clustering.

6.4 Schematic illustration of your method

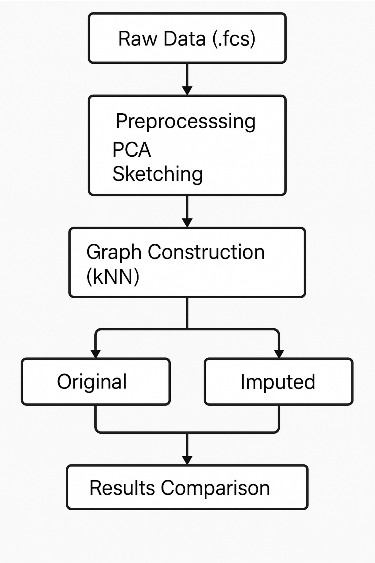


Figure 1: Overview of the Leiden clustering pipeline. The CyTOF data undergoes dimensionality reduction via PCA, followed by geometric sketching to sample representative cells. A k-nearest neighbors graph is constructed from the sketch, and Leiden clustering is then applied to identify immune cell communities. The same pipeline is applied to both original and imputed datasets.

7 Results

Table 2 and Table 1 show ΔCV for each biomarker in ascending order of absolute value (with labeling columns, such as barcodes, not included). A lower absolute value implies less deviation from the original dataset. We observe that the model trained on twelve patients for day 2 has overall lower variations in ΔCV, where the model trained on only two patients for day 14 has significantly higher variations. We conclude that EnGen, while a powerful tool, is less than ideal for small patient sample sizes due to a lack of variety in the training data.

As an example of reasonable and less-reasonable outputs from our feature imputation, we compare two plots of different biomarkers on each access from the day 2 model. Figure 3 shows two biomarkers with low ΔCV (164Dy_IkB, 139La_CD66). We observe that the imputed data, while maybe not completely representative of the original distribution in magnitude, does not vary wildly in its standard deviation and mean. Figure 2 shows two biomarkers with high ΔCV (172Yb_IgM, 148Nd_CD123), where the mean for the former is significantly shifted right on the x-axis and not representative of the original dataset at all.

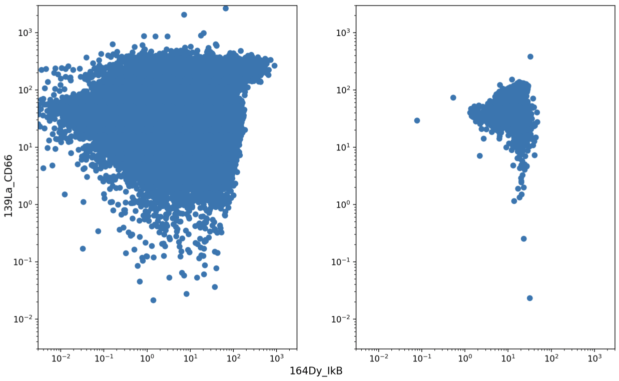


Figure 2: Example of two biomarkers with low ΔCV. The original data sample is on the left, and the imputed data sample (not concatenated) is on the right.

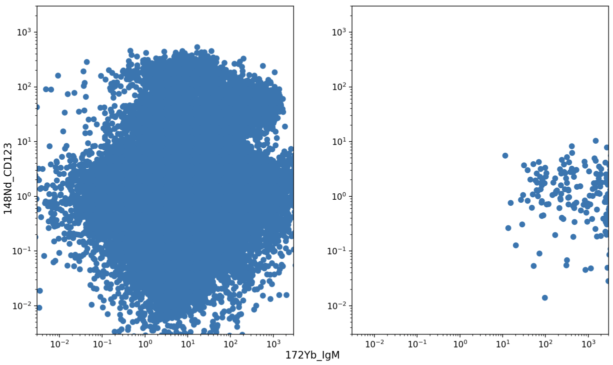


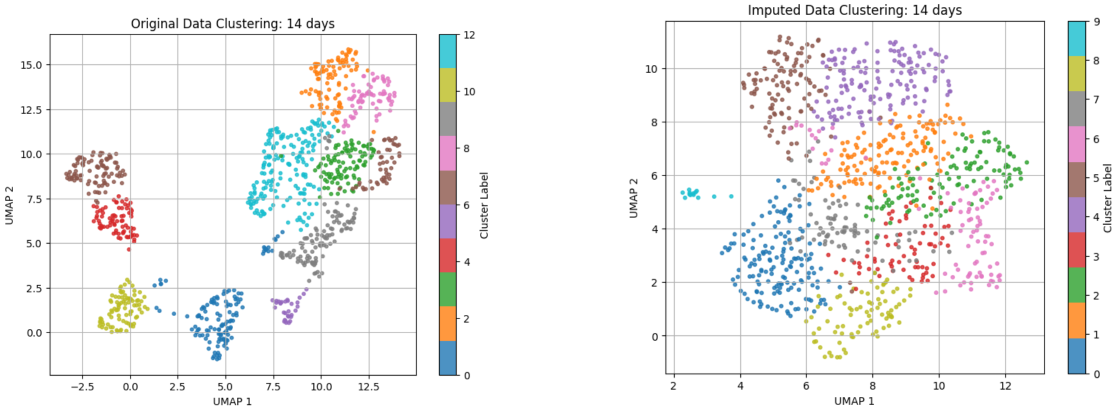
Figure 3: Example of two biomarkers with high ΔCV. The original data sample is on the left, and the imputed data sample (not concatenated) is on the right.

Label	ΔCV (Day 14)
164Dy_IkB	−30.68
139La_CD66	−48.19
166Er_NFkB	−49.20
143Nd_CD45RA	51.84
149Sm_CREB	−55.95
142Nd_CD19	57.92
157Gd_CD38	−63.28
115In_CD45	64.92
160Gd_Tbet	−84.55
156Gd_CD24	−96.85
144Nd_CD11b	−98.11
176Lu_CD56	−104.29
168Er_pSTAT6	−120.76
150Nd_STAT5	−132.21
165Ho_CD16	−144.10
154Sm_STAT3	−147.74
162Dy_FoxP3	−151.10
147Sm_CD11c	−154.42
171Yb_CD27	−160.43
159Tb_MAPKAPK2	−162.34
158Gd_CD33	−190.30
170Er_CD3	200.63
173Yb_CCR2	−210.77
175Lu_CD14	−213.50
141Pr_CD7	235.85
151Eu_p38	−239.66
167Er_ERK	−240.09
155Gd_S6	−270.55
174Yb_HLADR	−285.36
169Tm_CD25	−299.59
113In_CD235ab_CD61	−309.44
145Nd_CD4	361.00
161Dy_cPARP	−371.70
153Eu_STAT1	−384.41
152Sm_TCRgd	−471.80
172Yb_IgM	−551.28
148Nd_CD123	1690.06
146Nd_CD8a	81575.41
Mean	2370.53
Median	161.38
# > 50	35

Table 1: Comparison of ΔCV across models trained to generate single-cell data at days 14. Each row contains a biomarker and its ΔCV for both models. The final three rows are the average, median, and number above 50% across all biomarkers.

Label	ΔCV (Day 2)
144Nd_CD11b	−0.21
167Er_ERK	7.88
115In_CD45	−14.72
162Dy_FoxP3	−14.86
149Sm_CREB	−15.09
171Yb_CD27	−17.34
150Nd_STAT5	19.87
158Gd_CD33	22.18
165Ho_CD16	−22.68
157Gd_CD38	−23.32
166Er_NFkB	−26.11
156Gd_CD24	−27.22
139La_CD66	−29.63
154Sm_STAT3	−38.27
164Dy_IkB	−43.03
155Gd_S6	−45.16
175Lu_CD14	53.44
143Nd_CD45RA	70.96
160Gd_Tbet	−72.28
176Lu_CD56	−76.92
168Er_pSTAT6	−106.68
172Yb_IgM	110.30
145Nd_CD4	−127.66
142Nd_CD19	−141.78
170Er_CD3	−144.78
141Pr_CD7	−147.93
153Eu_STAT1	−149.00
146Nd_CD8a	−205.53
148Nd_CD123	−224.67
151Eu_p38	252.86
147Sm_CD11c	259.11
173Yb_CCR2	346.21
159Tb_MAPKAPK2	394.95
113In_CD235ab_CD61	−492.60
161Dy_cPARP	2980.76
169Tm_CD25	5650.61
152Sm_TCRgd	12363.99
174Yb_HLADR	34181.50
Mean	1550.58
Median	74.60
# > 50	22

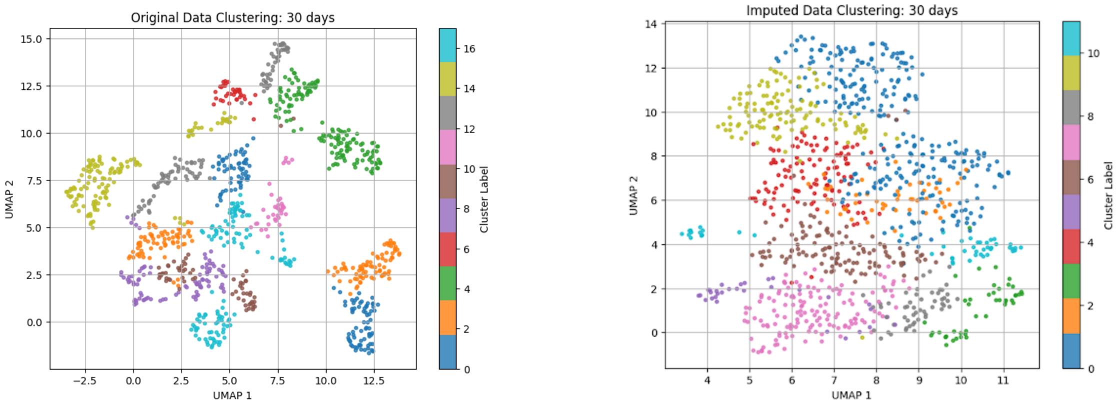
Table 2: Comparison of ΔCV across models trained to generate single-cell data at day 2. Each row contains a biomarker and its ΔCV for both models. The final three rows are the average, median, and number above 50% across all biomarkers.



(a) original data

(b) imputed data

Figure 4: UMAP representation from Leiden clustering on original (left) vs. imputed (right) features for day 14



(a) original data

(b) imputed data

Figure 5: UMAP representation from Leiden clustering on original (left) vs. imputed (right) features for day 30

Moreover, we show results from Leiden clustering on the original vs. imputed features of the dataset for each day we tried imputing features. Figure 4 and Figure 5 show UMAP visualizations of both. Figure 6 shows the number of communities detected for each time point in the Leiden clustering on the original dataset, and Figure 7 shows UMAP visualizations for each day in the original study's data.

7.1 Datasets

We used one dataset for this project. The dataset is the same as presented in the post-stroke immune system profiling study we aimed to recreate. This dataset contained many FlowCytometry files in .fcs format split by patient and timepoint in the study. Each file contained approximately six-hundred thousand cell samples using a 47-parameter assay (including labeling columns such as barcodes). Additionally, we would like to thank Dr. Stanley for making this dataset easily accessible to our group for the purposes of this project.

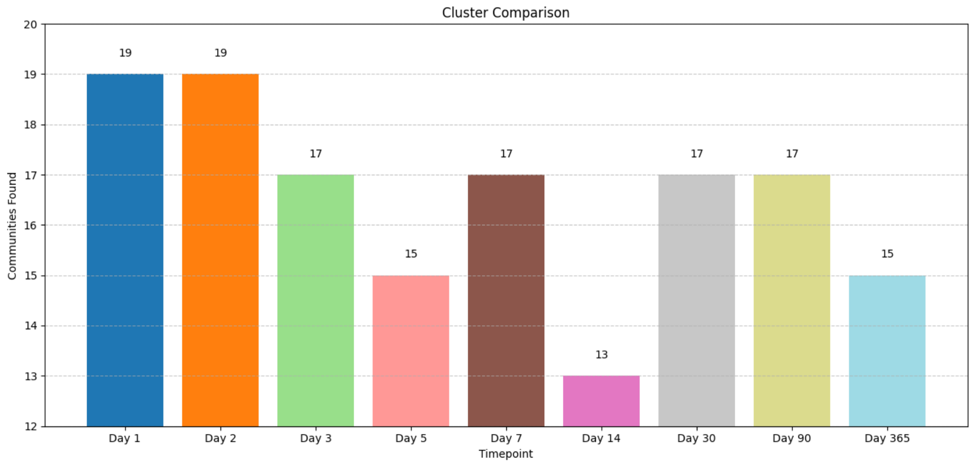
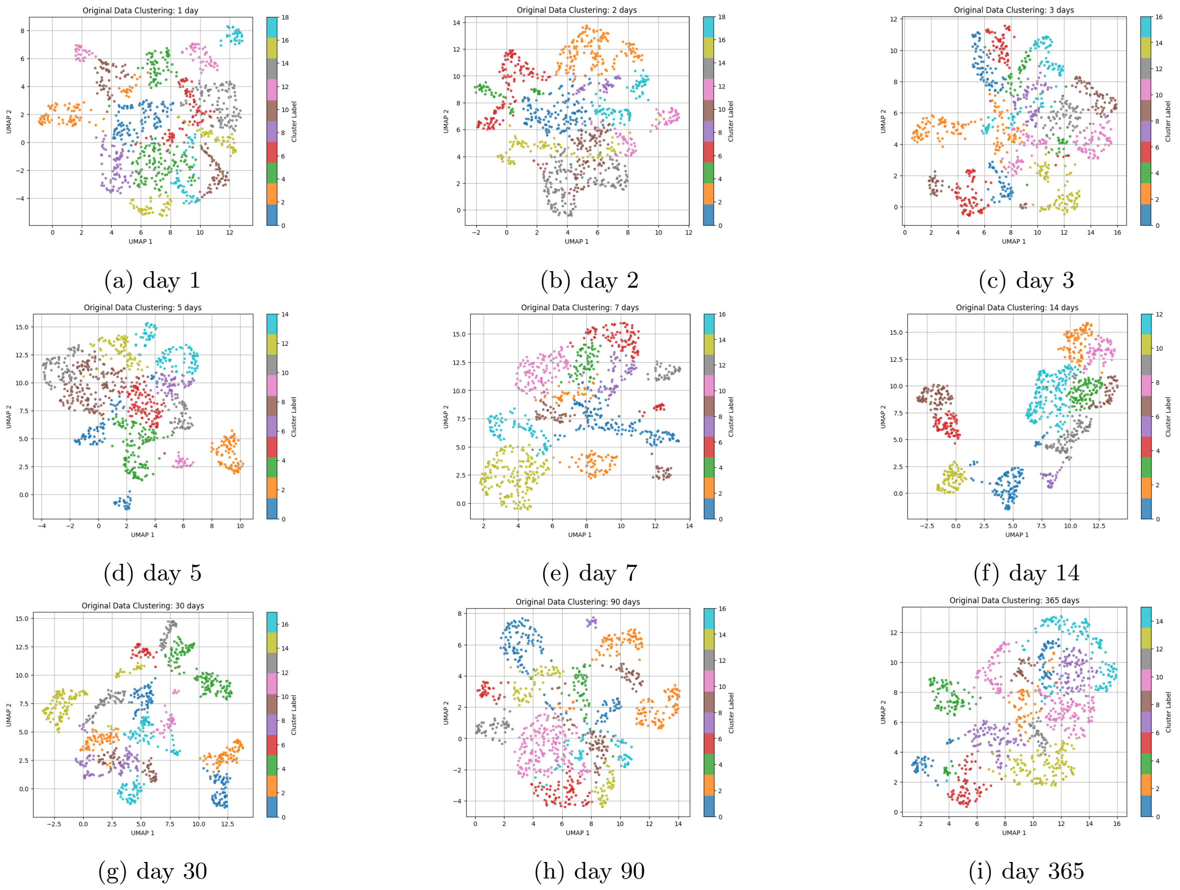


Figure 6: Number of clusters per time point in the dataset.



(a) day 1

(b) day 2

(c) day 3

(d) day 5

(e) day 7

(f) day 14

(g) day 30

(h) day 90

(i) day 365

Figure 7: UMAP representation from Leiden clustering on all time points of data.

7.2 Baselines

For a baseline of feature imputation, we trained two additional models on timepoints 1 and 2 using the first twelve (approximately half) patients' samples and timepoints 1 and 14 using two (patients 2 and 3) of the patients' samples, where we validated the quality of generations on the 13th patient and 12th patient, respectively. This approach of training on a larger cohort of patients and testing on one patient is consistent with the approach demonstrated in the original EnGen paper.

7.3 Description of Experiments

We compared the generated data for each held-out patient to their actual single-cell data by measuring the delta coefficient of variation between their actual samples and their samples combined concatenated with their imputed features. We compare the original samples coefficient of variation to the concatenation of original and imputed features to make a true apples to apples comparison between how much the imputation process affects the distribution of features from the samples, because directly comparing the coefficient of variation between the original and imputed features alone could theoretically provide false assurances in data quality. We showed that while the baseline imputed data is not close to a perfect generative model for patient single-cell data, it is significantly better than a model trained on only two patients. We note that due to computational limitations, we only trained each EnGen model for 1000 epochs compared to the original implementation's 2500 epochs.

When performing Leiden graph clustering on the original patient data, days 1, 2, 3, 7, 30, and 90 produced between 17 and 19 communities each. Days 5 and 365 produced 15 communities each. Day 14 produced 13 communities.

When performing Leiden graph clustering on the imputed patient data, day 14 produced 10 communities and day 30 produced 11 communities. To compare how Leiden clustering on the original data compared to imputed data, we calculated a Normalized Mutual Information (NMI) score that compares the cluster labelings. Day 14 had an NMI score of 0.0340, and Day 30 had an NMI score of 0.0528.

8 Discussion

8.1 Recap

In summary, we focused our efforts on feature imputation on the two time points in the original study that had the least number of patients due to computational constraints. We utilized EnGen, an encoder-decoder deep neural network specialized in generating single-cell data across time points, to attempt feature imputation for these days with heavy patient dropout. We analyzed the results of our feature imputation process, by comparing one model (day 14) against a model trained with significantly more patient's samples (day 2), and demonstrated a lack of quality data imputation for the former due to low patient availability at that time in the study. Additionally, we performed Leiden clustering on geometric sketches of all patient data from each time point in the original study. We showed distinct clusters forming from the original data that were consistent with the number of communities found in the original paper. We further compared these results to Leiden clustering on the imputed data for days 14 and 30, and we showed a lack of distinctive clusters at those time points.

8.2 Observations

Previously, 18 communities corresponding to immune features, which changed together as patients recovered from a cerebrovascular accident (stroke), were identified through a correlation network (Tsai et al. 2019). These communities played an important role in identifying the three post-stroke immune system response phases in Tsai et al. 2019. Since our goal was to investigate how data imputation may impact the number and nature of immune system response phases that were previously identified on the limited patient data, we decided to search for communities in a different manner. To determine how graph clustering approaches would compare to their correlation network, we started by employing Leiden clustering (Traag et al. 2019) on the original CyTOF patient data for all timepoints. Since some time points included over 600k cells, we first did a geometric sketch (Hie et al. 2019) to create a representative subset of the patient data while reducing computational demands. Then, we performed Leiden clustering to identify unique cell communities. Leiden clustering resulted in 17 to 19 communities for six of the nine blood-collection timepoints. These community counts closely mirrored the 18 communities identified by the correlation network in Tsai et al, 2019, indicating that this graph representation of the original patient data may be useful for identifying post-stroke immune features. Interestingly, the patient data for Day 14 post-stroke produced only 13 communities in Leiden clustering. Since Day 14 also had the lowest number of patient samples (3 out of 25 patients gave blood on this day), we decided that Day 14 would be a prime candidate for data imputation, as generating synthetic data for this time point may uncover hidden biomarkers.

The Leiden clustering results on the imputed datasets for Days 14 and 30 were unexpected. Rather than revealing a graph clustering structure similar to the original day 14 and 30 data, or at least similar to other data-collection timepoints, the imputed data Leiden clustering appeared to have far lower cluster separation, based on our observations of the 2D UMAP visualizations. We define cluster separation as how well-isolated the different groups of cells appear in the embedding space. A higher cluster separation tends to reflect how well the clustering has captured biological differences in the cell population, so this imputed data appears to have lower biological heterogeneity than the original data. Given that the original data from time points that had nearly all 25 patients give blood have much greater cluster separation than these imputed datasets, these imputed data may be nonfunctional. Additionally, day 14 imputed data showed just 10 clusters compared to the original 13 clusters, and day 30 imputed data showed just 11 clusters compared to the original 15 clusters. The data imputation was expected to increase the community count on day 14, but the results ran counter to this expectation. NMI scores of 0.034 and 0.053 for days 14 and 30, respectively, indicate very poor cluster alignment.

8.3 Limitations and Future Work

Limitations. While the Leiden clustering pipeline using PCA and geometric sketching is computationally efficient and effective for large-scale data, it comes with several limitations. First, reducing the dataset to 20 principal components via PCA inevitably results in information loss. Important nonlinear relationships or signals from rare cell types might be discarded in the process, potentially affecting the ability to distinguish subtle biological differences. Second, geometric sketching introduces a sampling bias by selecting only a subset of representative cells—typically around 1,000. This method may overlook rare populations, especially in imbalanced datasets, which can lead to an underestimation of cluster diversity. Third, Leiden clustering is sensitive to the resolution parameter. While a fixed resolution value (e.g., 1.0) is commonly used, it may not be optimal for every dataset, possibly causing over- or under-clustering depending on the data structure. Finally, when imputed data is used, the risk of introducing artifacts increases. Imputed values may reflect model-based assumptions rather than true biological signals, leading to the formation of artificial or misleading clusters. Together, these limitations suggest that while the Leiden clustering approach is powerful, careful parameter tuning and validation are essential to ensure biologically meaningful results.

Future Work. Due to the limited access to the GPUs, we could not train the EnGen model as recommended in the EnGen paper. In the original paper, they trained on GPUs with a total of 2500 epochs. For future work, with access to the GPUs, we can allow the model to train longer, hence improving the reconstruction quality. While this may not drastically improve feature imputation due to the very small samples available at days 14 and 30, it would likely make some amount of improvement. Additionally, we would like to attempt other, more computationally efficient methods for imputing single-cell features like GANs. Finally, given the change in MoCA scores for each patient, we would like to complete this project by training a classifier model to predict which principal components of each cluster assignment from each time point, compared to day 365, have a clinically predictive capability. This would be the final piece of data necessary to complete the recreation of the original study with a clustering-based approach.

8.4 Inspiring Concluding Paragraph

Although the feature imputation and graph clustering approach to post-stroke immune system profiling did not yield the results we initially hoped for, we absolutely learned a great deal about computational biology in the process. Clinical studies, in particular, are not particularly easy to recreate strong results for due to a lack of data from patient dropout. However, that does not mean that we should not investigate this problem further, as patient dropout will continue to be an issue for clinical studies for the foreseeable future. The knowledge we gained from the complexities of single-cell data, ways to analyze feature imputation outputs, and re-applying concepts learned from our early days in class galvanizes us to continue exploring graph-based approaches in complex data domains.

9 References

- Fallahzadeh, Ramin, Neda H. Bidoki, Ina A. Stelzer, Martin Becker, Ivana Marić, Alan L. Chang, Anthony Culos et al. ”In-silico generation of high-dimensional immune response data in patients using a deep neural network.” Cytometry Part A 103, no. 5 (2023): 392-404.

- Tsai, Amy S., Kacey Berry, Maxime M. Beneyto, Dyani Gaudilliere, Edward A. Ganio, Anthony Culos, Mohammad S. Ghaemi et al. ”A year-long immune profile of the systemic response in acute stroke survivors.” Brain 142, no. 4 (2019): 978-991.

- Traag, V. A., L. Waltman, and N. J. Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. 9, 5233. 2019. Pendlebury, Sarah T., and Peter M. Rothwell. ”Prevalence, incidence, and factors associated with pre-stroke and post-stroke dementia: a systematic review and meta-analysis.” The Lancet Neurology 8, no. 11 (2009): 1006-1018.

- Enders, Craig K. Applied missing data analysis. Guilford Publications, 2022. Zou, Hui, and Trevor Hastie. ”Regularization and variable selection via the elastic net.” Journal of the Royal Statistical Society Series B: Statistical Methodology 67, no. 2 (2005): 301-320.