Bioinformatics, 2024, 1–16

Density-Based Clustering in Immune Cell Cytometry Data

Rishika Madhanagopal

Bioinformatics, Newcastle University, Newcastle Upon Tyne, United Kingdom

Corresponding author. R.Madhanagopal2@newcastle.ac.uk

Abstract

Recent advancements in flow cytometry have revolutionized the study of the immune system, enabling detailed single-cell analysis that provides valuable insights into immune-mediated diseases such as rheumatoid arthritis (RA). This study leverages data from the BioFlare project, a longitudinal cohort study focusing on the pathogenesis of RA flares, to apply density-based clustering, to flow cytometry data. The aim is to identify distinct immune cell populations and potential biomarkers associated with disease flares and remission. The dataset, comprising immune cell profiles from RA patients, was pre-processed and subjected to OPTICS clustering, followed by visualization using UMAP and t-SNE techniques. While the study revealed the potential of density-based clustering, challenges in obtaining well-defined clusters due to the complex, high-dimensional nature of the data were evident. Differential analysis could not be completed due to these limitations. Despite this, the research provides important insights into the application of advanced computational techniques in immunology and suggests pathways for future work, including refining clustering processes and validating identified clusters to enhance the understanding of RA and support personalized treatment strategies.

Key words: Rheumatoid Arthritis (RA), Flow Cytometry, Density-Based Clustering, BioFlare Project,Immune Cell Populations, Biomarkers, UMAP, t-SNE,Immunology,Personalized Treatment Strategies

1 Introduction

In recent years, the study of the immune system has been revolutionized by advancements in flow cytometry, enabling detailed analysis of individual cells and offering unprecedented insights into the immune system's complexity. This technique allows researchers to measure the proportion and characteristics of different immune cell populations within patients, providing a comprehensive snapshot of the immune landscape. Such detailed profiling opens new avenues for understanding the mechanisms underlying immune-mediated diseases, such as **rheumatoid arthritis (RA)**, and offers potential pathways toward more personalized treatment strategies.

Despite these advantages, the application of **Density-Based Clustering Algorithm** to RA cytometry data remains relatively unexplored. The Bioflare project, a prospective longitudinal cohort study funded by the Medical Research Council (UK) and led by Newcastle University in collaboration with the University of Glasgow and the University of Birmingham, offers a unique opportunity for rectifying this. BIO-FLARE stands for BIOlogical Factors that Limit sustained Remission in rheumatoid arthritis. The research is focused on the pathogenesis of RA flares and aims to identify associated biomarkers. In this respect, it is based on an extensive dataset drawn from Bioflare containing flow cytometry profiles of immune cells in RA patient's flare and non-flare conditions. This dataset provides a paramount opportunity for the application of density-based clustering in the identification of distinct populations of immune cells and could also identify novel biomarkers of RA flares. The present study, by use of this data, proposes the determination if density-based clustering identifies cell populations and biomarkers with regard to clinic relevance for a better understanding of RA and improved treatment strategies [**1**].

The aim of this research is to evaluate the effectiveness of density-based clustering, specifically **OPTICS** (Ordering Points to Identify the Clustering Structure), in identifying distinct immune cell populations and potential biomarkers from pre-processed cytometry data. The study will also explore how these identified clusters and biomarkers relate to clinical outcomes, particularly in distinguishing between patients experiencing flares and those who are not. Additionally, the research seeks to develop visualizations to aid in the interpretation of these findings and, if feasible, perform differential analysis to further elucidate the clinical significance of the identified clusters and biomarkers.

Rheumatoid arthritis is a chronic, systemic autoimmune disease affecting about **1%** of the world population, the majority being women. The illness is characterized by chronic synovitis leading to pain, swelling, and finally destruction of joints in the absence of treatment [**1**] [**2**] [**3**]. While RA itself is not clearly defined regarding its origin, it is considered caused by genetic predisposition, environmental factors, and immune dysregulation. Probably greatest challenge to the management of RA lies in its heterogeneity itself, very variable, often complex symptomatology and courses of disease or different responses to treatment in patients, thus making the selection of potential biomarkers with regard to predicting disease activity, flares, or remission and guiding treatment decisions very hard [**4**] [**5**] [**6**].

Flow cytometry has become one of the great investigative tools in immunology, enabling detailed profiling of immune cell populations in diseases like RA. Such is the case for flow cytometry, it measures multiple parameters at the single-cell level, generating high-dimensional data that describe the constitution and function of the immune system. However, traditional methods for analysis, which use manual gating, are limited only by their intrinsic subjectivity and inability to deal with such complexity in these data [**7**]. This has led to investigations of automated clustering algorithms that may allow more objective identification of distinct cell populations within the data. It is within these algorithms that density-based methods, such as **DBSCAN** and **OPTICS**, previously showed potential for identifying clusters of varying densities and shapes, making them very applicable to the analysis of complex biological data like flow cytometry [**31**] [**9**] [**10**].

The primary objective of this study is to first apply the **OPTICS** algorithm to pre-processed flow cytometry data from RA patients, aiming to identify distinct immune cell populations and potential biomarkers associated with disease flares and remission. Second, to conduct **differential analysis** on the identified clusters to determine which biomarkers are significantly different in flare conditions compared to non-flare states in RA patients, thereby contributing to the understanding of disease activity. Finally, the study aims to provide **insight into the pathogenesis** of RA flares by correlating the identified immune cell populations with clinical outcomes and significant biomarkers, which could enhance our understanding of RA progression and inform potential treatment strategies. Unlike traditional clustering techniques, which require a predefined number of clusters [**11**] [**12**], OPTICS can uncover the intrinsic clustering structure of a dataset, making it highly suitable for exploratory analysis. Additionally, dimensionality reduction techniques like UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-Distributed Stochastic Neighbor Embedding) will be employed for intuitive visualization of the clustering results, enhancing the interpretability of these complex data sets [**13**] [**14**].

Briefly, this research seeks to advance our understanding of RA pathogenesis by employing state-of-the-art unsupervised clustering and visualization tools. The insights gained may lead to the identification of novel biomarkers, ultimately contributing to more personalized treatment strategies and a better understanding of disease activity in RA.

2 Background

Rheumatoid Arthritis (RA) is a complex and chronic autoimmune disease that affects millions worldwide, primarily targeting the joints and leading to significant morbidity. Understanding RA is crucial not only because of its prevalence but also due to the substantial challenges it poses in terms of treatment and disease management. The primary challenge lies in the disease's heterogeneity and the difficulty in identifying the factors that drive its relapsing-remitting nature. In this section, we will explore what RA is, why it is worth studying, the challenges involved in its treatment, and how advances in computational techniques, particularly in flow cytometry data analysis, can contribute to better understanding and managing this condition.

The immune system comprises a complex network of cells, tissues, and organs whose main duty is the protection of the body against potentially pathogenic organisms [**15**]. Given the complexity of the system, especially the different populations of immune cells and their interactions, this understanding is very important in the management and treatment of immune-mediated diseases like rheumatoid arthritis.

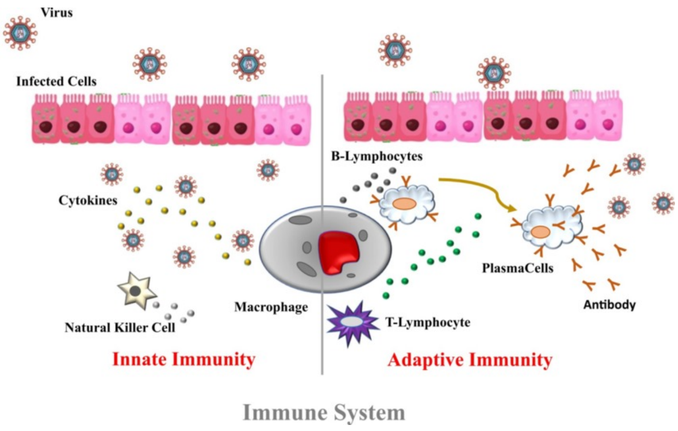

Fig. 1: *Mechanism illustrating how immune system works* [**18**]

Rheumatoid Arthritis (RA) is a chronic autoimmune disease, primarily affecting the joints, characterized by persistent synovitis, leading to pain, swelling, and eventually, if not treated properly, progressive destruction of the joints. RA affects an estimated **0.5-1%** of the world's population, predominantly females between the ages of 30 and 60 [**2**] [**3**] [**16**]. RA is characterized by the presence of autoantibodies, such as **rheumatoid factor** and **anti-citrullinated protein antibodies** with systemic manifestations include major cardiovascular disease, lung pathology, anemia, and other complications beyond joint involvement that much impair the quality of life of these patients [**4**].

RA belongs to the large category of disorders known as **Immune-Mediated Inflammatory Diseases** (IMID), which are characterized by dysregulated immune responses that finally lead to chronic inflammation. The inflammation in IMIDs, like RA, follows a relapsing-remitting course; at periodic intervals, a phase of active disease is followed by remission. Very little is understood about what factors are responsible for this alternation between active and quiescent states, and it becomes a challenge to the management and treatment of diseases [**17**].

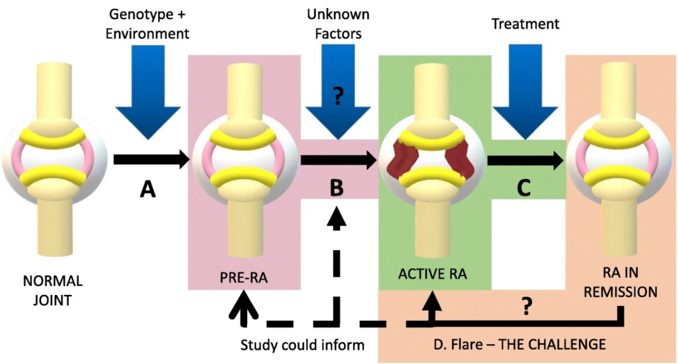

Fig. 2: *RA Aetiopathogenesis* [**1**]

The pathogenesis in RA results from a complex interplay between genetic predisposition, environmental factors, and immune system dysregulation. The immune system, which normally guards the body against infections in other cases, misbehaves and becomes hostile toward the synovial tissue in the joints of the body, causing perpetual inflammatory processes and subsequent damage to joints. RA involves populations of very heterogeneous immune cells, including T cells, B cells, macrophages, and dendritic cells, each uniquely contributing to disease progression. For example, the recognition of self-antigens by T cells is essential to induce RA and to activate other immune cells; B cells classically produce autoantibodies that drive chronic inflammation in RA [4] [**19**] [**20**] [**21**].

The study of the composition and behaviours of these two immune cell populations is relevant to the development of effective treatments for RA, in particular to the identification of biomarkers predictive of disease activity that may guide therapeutic decisions.

Flow cytometry is the technique in use to date, which permits the evaluation of physical and chemical characteristics of cells. It allows for the performance of multiple, diverse parameters in thousands of single cells per second, thus giving full profiles of immune cell subsets based on surface markers and intracellular cytokines, among other cellular features. Within RA, flow cytometry shapes immune cell populations and aids in the identification of biomarkers for the characterization of the immunopathogenesis of this disease . One of the biggest challenges in the analysis of flow cytometry data is its high dimensionality. Traditionally, analysis has been performed through the visual inspection of data to define cell populations, often by **manual gating**. This approach is subjective, extremely time-consuming, and may miss rare or subtle populations. In turn, interest is growing in **automated clustering algorithms** that allow for a more objective, hence more scalable, way of analysing cytometry data with the potential to yield novel insights into immune cell behaviour [**7**].

Clustering is one of the most important tasks in data analysis, which aims to find natural groupings in a dataset based on similarities among data points. Traditional clustering algorithms, such as K-means, have pre-fixed the number of clusters to choose from and often make an assumption about spherical cluster shapes [**12**]. In contrast, most real biological and medical datasets are known to include clusters of different shapes, sizes, and densities, hence requiring more advanced clustering techniques that can discover such structures without any predefined assumptions [**11**].

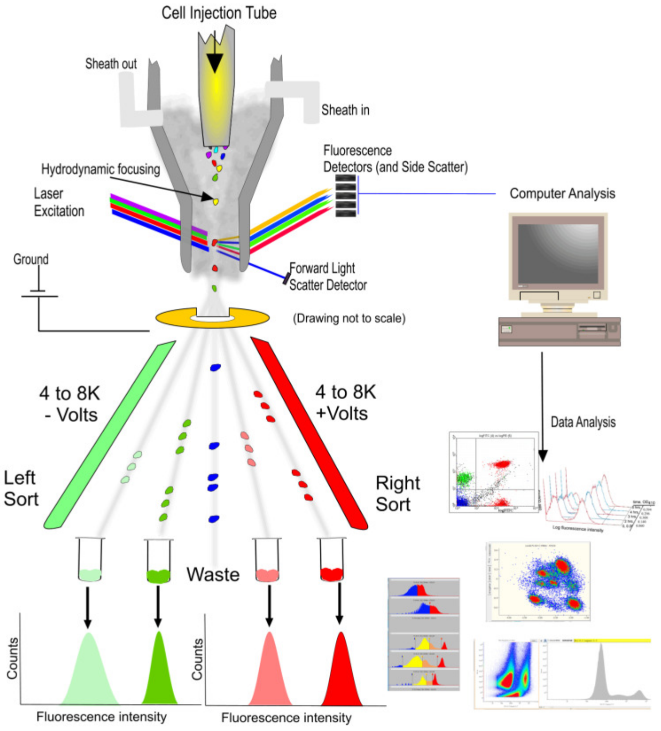

Fig. 3: *General scheme of a flow cytometer showing isolation of single cells* [**22**]

Among density-based algorithms of clustering, two of them are appropriate for the analysis of complex biological data: **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) and **OPTICS** (Ordering Points to Identify the Clustering Structure). These algorithms define clusters as regions of high point density separated by regions of low point density. The OPTICS, proposed by **Ankerst et al.** in **1999**, construct a very robust density-based algorithm for clustering, designed in order to overcome the weaknesses of the preceding algorithms, especially DBSCAN. OPTICS ranks the points according to their density and identifies clusters of different densities, all without the use of any predefined number of clusters. This makes it a flexible and powerful tool while analysing such complex data like flow cytometry [**31**].

Despite these advantages, the application of OPTICS to RA cytometry data remains relatively unexplored. The Bioflare project, a prospective longitudinal cohort study funded by the Medical Research Council (UK) and led by Newcastle University in collaboration with the University of Glasgow and the University of Birmingham, provides an ideal opportunity to address this gap. The project focuses on the pathogenesis of RA flares and aims to identify associated biomarkers using an extensive dataset from RA patients during flare and remission conditions. This dataset offers a unique opportunity to apply OPTICS clustering to identify distinct immune cell populations and novel biomarkers related to RA flares.

This study aims to systematically investigate the efficiency of OPTICS clustering in RA cytometry data and its potential for identifying clinically relevant biomarkers. Additionally, the study seeks to explore how the identified clusters correlate with clinical outcomes, such as disease flares or remission, ultimately contributing to personalized treatment strategies for RA patients [**1**].

3 Literature Review

3.1 Cytometry Data Analysis

Flow cytometry has been one of the most principal techniques in immunology for decades and allows one to conduct detailed immune cell profiling based on multiple parameters at the single-cell level. That is one important information in understanding the complexity of the immune system, particularly in autoimmune diseases such as rheumatoid arthritis [**23**] [**24**]. Traditional methods of looking at this data have been manual gating, and these techniques have been used by many people as these are very simple and easy to directly apply. However, the increasing scale and complexity of cytometry data are raising demands for advanced analytical techniques.

Although manual gating is the most widely used methodology, it is currently subject to many criticisms [**12**]. It is very subjective and often results in irreproducible results even for the same experienced analyst [**25**]. An increase in the optimal number of events would result in increased false positive due to the inability of manual gating to perfectly find the major subsets when they are numerous. Also importantly, manual gating is very time-consuming, a feature that continues to pose real challenges to researchers as the number of parameters and hence the number of cell populations that need to be analyzed grows [**7**]. However, the above-listed drawbacks of the manual gating process encourage the development of high-level computational methodologies to achieve consistent and robust identification and annotation of cell populations within flow cytometry data, hence eliminating subjective and non-scalable aspects of the traditional processes [**26**].

3.2 Clustering Analysis

In view of this, high interest in applying machine learning and data mining techniques to analyze cytometry data has been exhibited. Much more promising to overcome the limits of manual gating are automatic clustering algorithms, specifically, density-based methods like OPTICS [**31**]. In particular, these algorithms work well with high-dimensional and complex characteristics of cytometry data, identifying cell clusters that differ in shape and density without the requirement for a predefined number of clusters [**11**]. This capacity is critical in diseases like RA, where heterogeneity of the immune system can obscure clinically meaningful small immune cell populations.

Furthermore, density-based clustering methods, like OPTICS, have been able to identify cell populations that are too rare or too subtle to be detected by traditional manual gating . In RA, such methods will be helpful in attempting to identify new biomarkers to distinguish the two phases of the disease: flare and remission. While such methods bring enormous advantages, they also present some significant challenges. For instance, true fluorescence still remains difficult to segregate from artefactual signals caused either by cellular autofluorescence or fluorescent spillover. At the same time, a lack of controls can lead to overestimation of cellular populations, which can easily complicate result interpretation through increased dimensionality of the dataset [**7**] [**27**].

3.3 Application to Other Immune-Mediated Diseases

Application of the advanced clustering algorithms is not restricted to RA alone: very similar approaches are under exploration in other immune-mediated diseases, such as **Systemic Lupus Erythematosus** (SLE) and even **Multiple**

Sclerosis (MS). In SLE, for instance, flow cytometry does characterize the immune cell landscape, finding significant heterogeneity among patients . Studies, by application of automatic clustering methods, have described **novel** immune-cell subsets that are associated with disease activities and further enabled a better understanding of the pathogenesis of the diseases and possible therapeutic targets [**28**]. Along similar lines, clustering algorithms to analyze cytometry data applied in MS were helpful in identifying **distinctly different** immune cell populations at different courses of the disease. It has been possible to get an insight into immune mechanisms supporting MS disease and underline potential biomarkers for disease progression [**29**] [**30**] . It has implications for the wider use of advanced computational techniques in the study of other immune-mediated diseases, where traditional methods of analysis methods often fall short.

3.4 Challenges and Gaps in Cytometry Data Analysis

Even as the incorporation of new computational techniques into cytometry data analysis has been a step in the right direction, it is of essence to also consider their limitations and pitfalls. Some of the strengths of the methods lie in handling high-dimensional complexity of data to reveal patterns and relationships, inaccessible through other conventional means. However, the success still fundamentally rests with the quality of the data and how appropriate the algorithms are used. For example, some density-based clustering algorithms, like OPTICS, can be very sensitive to some parameters, such as the minimum number of points needed to form a cluster or the distance measure used. Thus, inappropriate choice of parameter leads to inappropriate clustering with either not including important clusters or identification of spurious ones. In addition, the algorithms can handle noise to a certain extent but not if it gets excessive in the data, then it will mislead the results [**31**]. Another important aspect is relating to proper interpretation of the results obtained from automated clustering methods. While these methods are able to identify new cell populations, their biological significance needs to be validated in experimental or clinical studies. Over-reliance on computational methods , without proper validation, may result in misleading conclusions. Thus, while advanced clustering algorithms afford powerful tools for the analysis of cytometry data, they should be combined with traditional methods and thoroughly tested for validation. Although density-based clustering techniques are gaining popularity in analyzing cytometry data, there are prominent gaps in our knowledge regarding the best ways of applying these methods in a clinical context. Specifically, very little is known about the performance of OPTICS clustering for identifying immune cell populations of clinical relevance in rheumatoid arthritis patients and how such findings might relate to clinical outcomes. Much of the literature has focused on more established methods, like k-means or hierarchical clustering, while the potential advantages gained through OPTICS have not been fully investigated. Furthermore, while dimensionality reduction using UMAP and t-SNE is used for visualization purposes, it is a rather new field in which to link its results to clustering for clinical interpretation. The aim of this study is to fill these gaps in knowledge by systematically investigating the efficiency of OPTICS clustering in RA cytometry data and its potential for identifying clinically relevant biomarkers. Another critical gap in understanding is how the clusters identified correlate with clinical outcomes like flares or remission. While some studies have implicated specific immune cells in disease activity, very little is known about which cell populations more closely relate to and predict clinical outcomes in RA. The current study therefore also seeks to fill this gap by using differential analysis to correlate identified clusters with clinical outcomes in an effort to eventually shed light on personalized treatment options for patients with RA.

4 Methods

The process of analysing the high dimensional data set I had was broken down into Three main steps, meticulously carried out with Python tools and libraries that made sure the analysis was

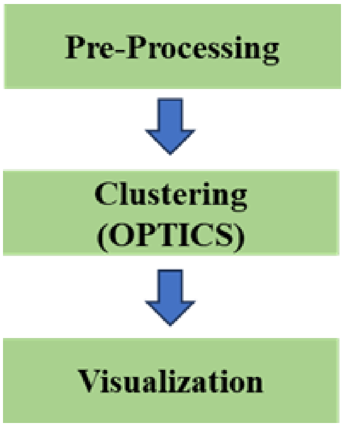

Fig. 4: *Steps involved in analysing high dimensional data*

strong enough to yield meaningful insights. 1) Pre-processing of the data, 2) clustering of the high dimensional cytometry data via OPTICS, 3) dimensionality reduction using UMAP and t-SNE for visualization purposes and differential analysis to identify unique cell populations.

4.1 Pre- Processing

4.1.1 Data Preparation

The dataset used in this study was obtained from the flow cytometry experiments run upon patients with rheumatoid arthritis. The data includes measurements of **16 biomarkers** across a large population of immune cells. The biomarkers are associated with various immune cell subsets, such as **CD4**, **CD8**, and **CD45**, which are critical in understanding the immunopathogenesis of RA. The dataset comprises hundreds of thousands of records, necessitating pre-processing before clustering. The data underwent essential pre-processing steps, including **transformation** using the **arcsine** function with a cofactor of **150** and **scaling** to a range of **0 to 1**. An arcsine transformation is particularly effective when the distribution of fluorescence intensity values in cytometry data is normally very skew. It helps to bring out the small values which might be overshadowed if such a transformation is not performed, and hence makes them comparable to the larger values for a more accurate analysis. The cofactor used is highly arbitrary in nature, but the use of this follows some of the common practices in the field and hence controls the sensitivity of the transformation so that the low and high intensity signals are scaled perfectly for analysis. Following transformation, scaling the data to a 0-1 range is critical for clustering algorithms, which require features to be on a comparable scale. Without scaling, features with larger numeric ranges could disproportionately influence the clustering process, leading to biased results. Therefore, these pre-processing steps are crucial to ensure that the data is normalized and uniform, making it suitable for reliable and meaningful analysis.

4.2 Clustering with OPTICS

4.2.1 Data Sampling

Starting this analysis, the data was read using **Pandas library**, a Python tool for flexible data processing. Due to the large size of the individual cell measurements dataset, which spans hundreds of thousands of records, a random **sampling** technique was employed to minimize computational load. More specifically, **10%** (approximately 50,000 records) was selected randomly from the original dataset by way of **Pandas sample method**, with a fixed random seed to ensure reproducibility of the results.

4.2.2 Implementation of OPTICS

The OPTICS algorithm was built with **sklearn.cluster.OPTICS** module from the **scikit-learn** library. To optimize clustering outcomes, multiple permutations of key parameters were tested systematically. The primary parameters explored were *xi, min_samples, max_eps, and min_cluster_size.*

4.2.3 Parameter Selection

In this regard, due to the complexity and variability of cytometry data, it was considered necessary to select suitable parameters for OPTICS. Nearly all clustering methods need values for input parameters that are difficult to determine, particularly in real-world datasets comprising high dimensional objects. The algorithms are quite sensitive to these parameter values, frequently generating very distinct partitioning of the dataset even with little parameter adjustments. In addition, high-dimensional real-datasets frequently feature

Table 1. Parameters for UMAP and t-SNE Methods

Parameter	Values Tested	Purpose
Xi	0.0001, 0.001, 0.01	Controls the steepness of the reachability plot, influencing cluster boundaries.
Min_samples	20, 50, 100	Minimum number of points to form a cluster, adjusted to identify small and large clusters.
Max_eps	0.5, 1.0, 2.0	Maximum distance between points in the same cluster, affecting sensitivity.
Min_cluster_size	5, 10, 20	Smallest allowable cluster size, crucial for detecting rare immune cell populations.

strongly skewed distributions that cannot be exposed by a clustering technique with only one global parameter value [**31**]. Initial parameter settings shown in **table 1** were chosen as an informed starting point based on prior knowledge of the expected size of immune cell populations. The points were then iteratively adjusted based on the **reachability plot**, a diagnostic tool that later aided in visualizing the clustering structure produced by OPTICS. In the present case, both **Euclidean** distance and **Manhattan** distance (metrics = "euclidean" or metrics = "manhattan") were tested to see which suits for high-dimensional data well and seems to capture the dissimilarities between points better in this context.

4.2.4 Rechability Plot

A reachability plot was a necessary aspect of the immune cell cytometry data analysis, especially for the visualization and understanding of the results from the OPTICS clustering algorithm. In this sense, after executing the OPTICS algorithm to the processed down sampled data, the reachability distances for each data point showed the order this algorithm had gone through [**31**]. The structure of the clusters tells valleys holding the aggregations of the immune cells, whereas peaks are noise. Most such valleys can be identified through their typical depth and width as distinct, statistically independent, and biologically relevant peaks, corresponding to different immune cell subpopulations in rheumatoid arthritis patients. Another important result that could be viewed from this reachability plot was the transitions between clusters; therefore, the data had a hierarchical nature and would allow to identify not only the primary clusters but also any smaller, nested clusters that might be present within them [**32**]. This visualization was very important in refining OPTICS parameters such as **xi, min_samples, max_eps, and min_cluster_size**, which optimize the formation of clusters while reducing noise. The reachability plot provided insights that allowed selection of robust clusters for further dimensionality reduction, therefore reflecting the underlying immune processes, thereby increasing the accuracy and relevance of results delivered in this study.

4.2.5 Tuning And Validation

The parameters have been iteratively tuned by running the OPTICS algorithm several times with different values and evaluating the the resulting clusters using the reachability plot. The goal was to maximize the number of meaningful clusters while minimizing noise.

4.3 Visualization

Following clustering, dimensionality reduction techniques were used to project down high-dimensional data to two dimensions for visualization. Two of the more popular techniques, **UMAP** and **t-SNE**, are used to create a 2D scatter plot of the clustered data.

4.3.1 UMAP (Uniform Manifold Approximation and Projection)

UMAP is an extremely powerful tool in dimensionality reduction and is recognized for its competitiveness with t-SNE in terms of visualization quality, with the added benefit of better preserving global structure and offering superior runtime performance [**33**]. This makes it eminently suitable for understanding the broader relationships between clusters. This approach was utilized in projecting high-dimensional cytometry data into a two-dimensional (2D) space. In this work, UMAP was implemented with the **umap-learn library**. Basically, UMAP works by creating a high-dimensional graph of the data and then optimizing a low-dimensional graph to capture the structure as much as possible from the original data. UMAP, on the other hand, is preferred since it balances global and local structure preservation, very often turning out to be important in the analysis of complex datasets with intricate patterns. By reducing data into a 2D space, UMAP will enable comparisons of clustering results from different parameter settings, providing insight into the most coherent and biologically relevant clusters.

4.3.2 t-SNE (t-Distributed Stochastic Neighbor Embedding)

To complement the global perspective provided by UMAP, t-SNE (t-Distributed Stochastic Neighbor Embedding), which is widely regarded as the state-of-the-art technique for dimensionality reduction in visualization [33], was used in parallel to provide an alternative visualization that emphasizes the local relationships within the data. t-SNE is especially known for its notion of revealing local structure by mapping similar data points close to one another in the 2D space [14]. T-SNE was implemented using the **sklearn.manifold.TSNE** module from the scikit-learn library.

Table 2. Parameters for UMAP and t-SNE Methods

Method	Parameters	Values
UMAP	n_components	2 (projected data into 2 dimensions for visualization)
	random_state	42 (ensures reproducibility)
t-SNE	n_components	2 (projected data into 2 dimensions for visualization)
	perplexity	30-50 (controls the balance between local and global aspects of data)
	n_iter	200-1000 (number of iterations for convergence)

Building upon the dimensionality reduction techniques, the 2D projections generated by UMAP and t-SNE were plotted using **Matplotlib**. For each algorithm, a scatter plot was created, where every point was colored according to the OPTICS-assigned cluster labels. That way, it would be easy to see how each one of the dimensionality reduction techniques would preserve the structure of the data and highlight different aspects of the clustering results. Scatter plots were qualitatively assessed for the distribution and separation of clusters. Quality was measured by compactness, separation, and known biological subsets of immune cells. If there were variations between the UMAP and t-SNE visualizations, the method providing the most biologically meaningful separation was prioritized for further interpretation.

5 Result

In this section, we present the outcomes of applying the OPTICS clustering algorithm to the pre-processed immune cell data from rheumatoid arthritis (RA) patients. The aim of this analysis was to identify distinct immune cell subpopulations by systematically varying the key parameters of the OPTICS algorithm. The results were evaluated in terms of cluster formation and reachability plots. We also explored the effectiveness of different dimensionality reduction techniques, including UMAP and t-SNE, to visualize the high-dimensional data in a two-dimensional space. The following subsections detail the clustering outcomes, the impact of parameter tuning, and the insights gained from the visualizations.

5.1 Clustering Outcomes with OPTICS and Reachability Plot Analysis

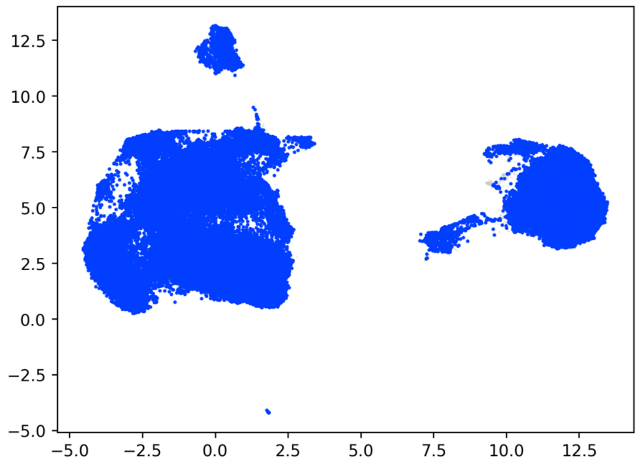

Fig. 5: *OPTICS clustering with Parameters: xi=0.01, min_samples=100, max_eps=1.0, min_cluster_size=10, showing a single large cluster across the dataset*

The above Scatter plot **Fig.5** shows the results from OPTICS Clustering algorithm with parameters

xi=0.01, min_samples=100, max_eps=1.0, and min_cluster_size=10. Most of the data points are colored the same, which means that this algorithm has grouped nearly the whole dataset into one cluster. This should point toward over-clustering, in which the algorithm misspoke in different meaningful clusters within the data. The lack of diversity in colors within clusters may mean the parameters are not very suitable to capture the real structure of the dataset at all. To address this, I analysed the reachability plots and identified the need for parameter tuning. . By adjusting these parameters, I aim to refine the clustering to reveal the underlying subpopulations that the current settings have overlooked.

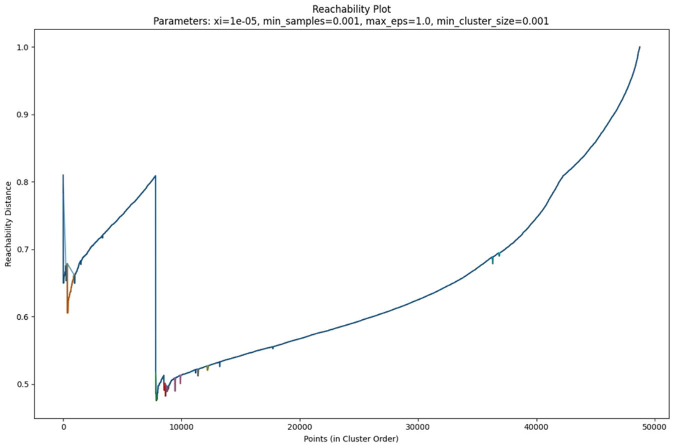

Fig. 6: *The plot shows several distinct valleys representing potential clusters and peaks indicating noise or outliers. The early part of the plot contains more defined clusters, while the latter part suggests more diffuse clustering and increased noise levels*

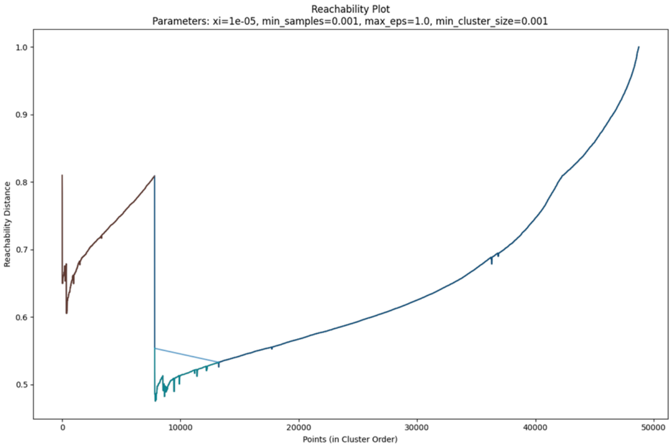

Fig. 7: *The plot identifies two main clusters (Cluster 0 and Cluster 1), with distinct valleys representing these clusters and a sharp rise indicating noise or outliers. Cluster 0 is well-defined with a deeper valley, while Cluster 1 is more diffuse, suggesting varying cluster densities within the data*

In the first reachability plot **Fig.6** generated by the OPTICS algorithm visually represents the clustering structure within the data. In this plot, the y-axis displays the reachability distance, while the x-axis is ordered by the points in cluster order. It has separate valleys and peaks in the figure; the valleys most probably show a possible cluster, and the peaks likely show noise or outliers. There are different colors in this plot corresponding to the various clusters obtained from OPTICS. What we can see is that the plot turns out to have a number of clusters, especially in the beginning part, with deep well-defined valleys indicating very clear and distinct groupings of data points. Later on in the plot, the valleys are shallow, indicating that some clusters are less distinct and more closely spaced, probably overlapping in feature space. Although peaks are very sharp early on, probably representing noise or outliers, the noise level does not appear to be very high. Conclusively, these results show that OPTICS has indeed found clusters with varying densities, some of them relatively poorly separated, so further tuning of parameters may help in drawing clearer distinctions between them. In contrast the second **Fig.7** reachability plot, shows us that there exist two distinct clusters, the rest of the points are classified as noise. Cluster, indicated by a deep valley in the early part of the plot, is more densely packed and well-defined, whereas Cluster appears as a shallower valley toward the middle, indicating that it is less dense and more diffuse. The steep climb of the reachability distance right after these valleys indicates the real transition to noise or outliers. This plot of the reachability distance against the algorithm while it processes the data illustrates the nicely separated clusters, with valleys indicating clusters and peaks showing noise. Already, the current parameters seem to capture the clustering structure quite well. However, the varying densities of the clusters and the number of clusters suggest there is space for further improvement of parameters toward a better separation of clusters or reduction of noise. The third reachability plot **Fig.8** (using the

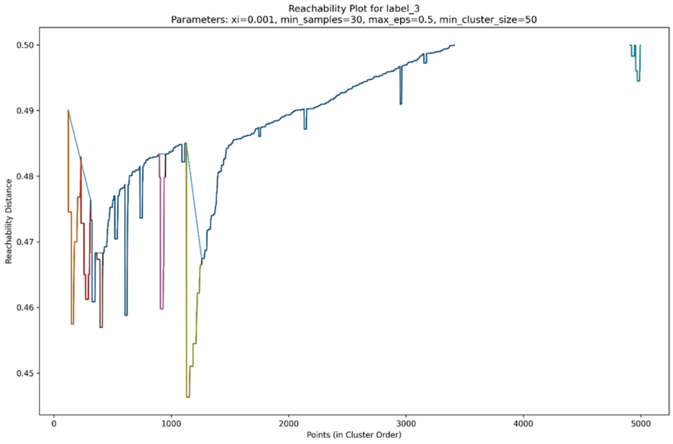

Fig. 8: *Rechability Plot With Manhattan Metric*

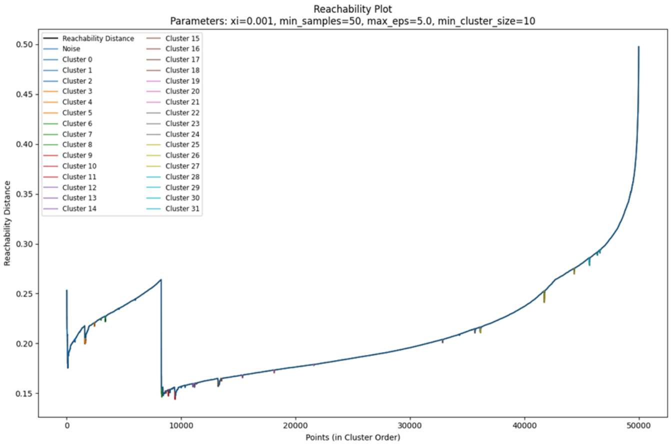

Fig. 9: *Rechability Plot With Euclidean Metric*

Manhattan distance metric) demonstrates several important observations which could be derived from the reachability plot with the **Manhattan distance metric**. The plot has various different valleys, most of them starting from the left, thus giving an indication of possible transitions between the segments of the data. The overall distances for reachability are very low, thus indicating compact groups within the dataset, varying from about 0.45 to 0.49. The right part of the plot with its gradual rise in reachability distance gives a hint about a broader, less defined segment or transition area. These valleys, by their size and depth, point to differences in the density or separation of data segments, some areas distinctly showing sharp distinctions from others. From such observations, one can characterize the existence of multiple and possibly distinct segments in a dataset with different internal densities and separations.

Finally ,from the fourth reachability plot **Fig.9**, it can be seen that there are 31 clusters as identified by the algorithm. However, shallow valleys and a smooth gradient in this plot suggest that the clusters might not be well-separated but overlapping or continuously distributed. Indeed, the lack of deep valleys usually characterizing strong, distinct clusters puts a question on the clarity and distinctiveness of the identified groups. This pattern might indicate over-segmentation as a result of the **Euclidean metric** of distance used during the clustering process, where there exist too many small clusters that actually are not strongly distinct groups.

5.2 Visualization with UMAP and t-SNE

5.2.1 Impact of Parameter Settings

The parameters used in the OPTICS algorithm, had a profound impact on the resulting clusters. For instance, when the min_cluster_size parameter was set higher (e.g., 50), the algorithm favoured the formation of larger, more robust clusters, as seen in **Fig.10**. Here, a significant portion of the data was left unassigned (represented as grey points), suggesting that the algorithm was stringent in forming clusters, potentially overlooking smaller or less dense groupings. Conversely, a lower min_samples setting (e.g., 30) in **Fig.11** allowed the algorithm to be more sensitive to smaller clusters, resulting in the identification of more clusters, though this also risked including noise or less distinct groupings. The max_eps parameter, controlling the maximum distance between points in a cluster, also influenced the results, with lower values leading to tighter, more compact clusters as shown in **Fig.12**.

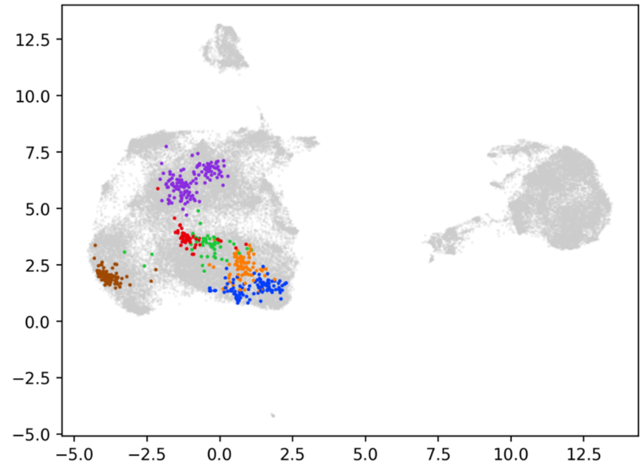

Fig. 10: *OPTICS clustering with xi=0.001, min_samples=30, and min_cluster_size=50 revealing 6 distinct clusters within the dataset.UMAP Projection with Manhattan Metric.*

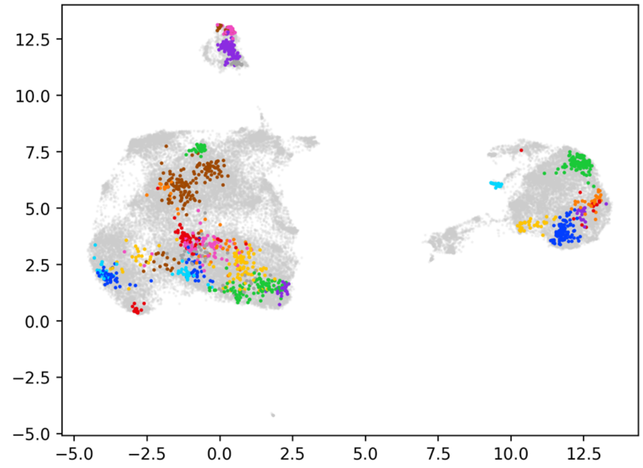

Fig. 11: *OPTICS clustering with xi=0.001, min_samples=30, and min_cluster_size=10 revealing 30 distinct clusters within the dataset.*

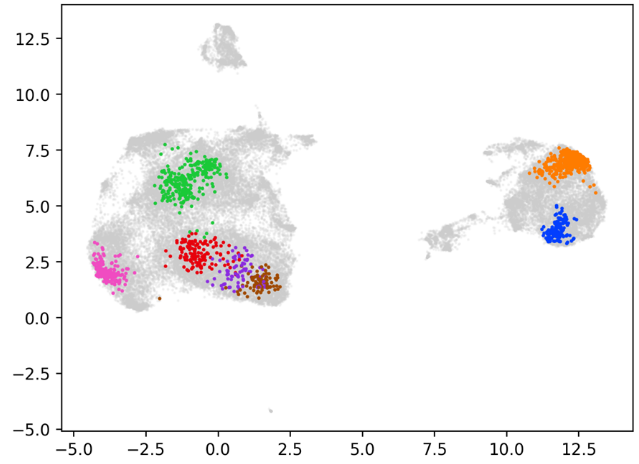

Fig. 12: *OPTICS clustering with xi=0.001, min_samples=100, max_eps=1.0 and min_cluster_size=50 revealing 7 distinct clusters within the dataset.*

5.2.2 Comparison of Distance Metrics

The choice between Manhattan and Euclidean distance metrics significantly impacted the clustering structure. The Manhattan metric produced more compact and distinct clusters, as seen in **Fig.10** (UMAP with Manhattan) and **Fig.13** (t-SNE with Manhattan). These metrics were particularly effective in the t-SNE visualization, where clusters were sharply separated with minimal overlap, indicating strong differentiation of data points. In contrast, the Euclidean metric, shown in **Fig.14** (UMAP with Euclidean) and **Fig.15** (t-SNE with Euclidean), provided a detailed view of the clustering structure but was more sensitive to internal variations within the data. This sensitivity led to finer distinctions between clusters, identifying subtle differences that might reflect rare populations or noise, as observed in the more granular distinctions and over-segmentation in these plots.

5.2.3 UMAP vs. t-SNE Visualizations

UMAP and t-SNE, both used for dimensionality reduction before visualizing the clusters, offered complementary perspectives. UMAP visualizations, especially when paired with the Manhattan metric **Fig.12**, provided a broader view of the clustering structure, preserving global relationships while still highlighting distinct clusters. This approach was useful for understanding the overall distribution of clusters across the dataset. On the other hand, t-SNE excelled at revealing local structure, focusing on the close relationships between points in the reduced space, as illustrated in **Fig.15** and **Fig.16**. The comparison shows that while UMAP offers a more holistic view, t-SNE is more effective at showcasing the fine details and separations within the data.

6 Disscussion

This study aimed to apply the OPTICS (Ordering Points to Identify the Clustering Structure) algorithm to flow cytometry data from rheumatoid arthritis (RA) patients to identify distinct immune cell populations and potential biomarkers associated with disease flares and

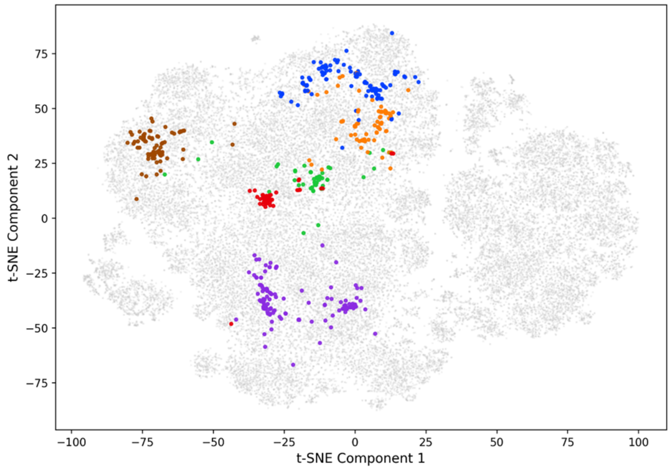

Fig. 13: *t-SNE Projection with xi=0.001, min_samples=30, min_cluster_size=50 and metric="manhattan" revealing 6 distinct clusters within the dataset.*

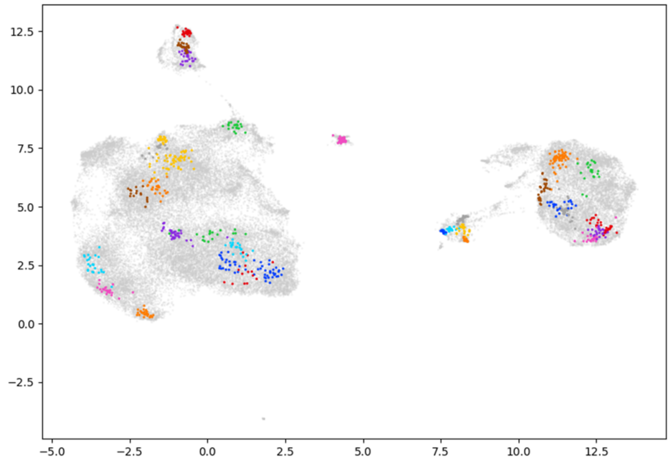

Fig. 14: *UMAP Projection with Min_sample 30, Euclidean Metric.*

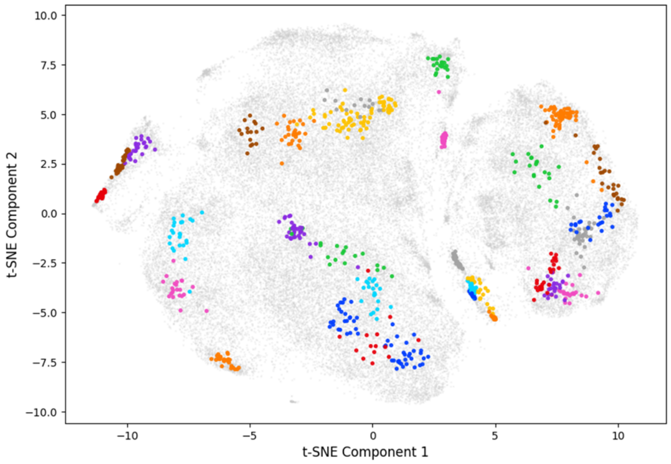

Fig. 15: *t-SNE With Euclidean Metric.*

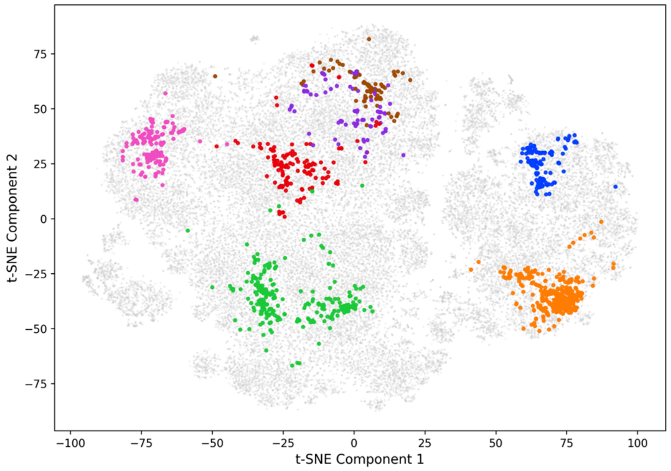

Fig. 16: *t-SNE Projection with xi=0.001, min_samples=100,metric="manhattan" and min_cluster_size=50 revealing 7 distinct clusters within the dataset.*

remission. The analysis involved preprocessing the data, implementing the OPTICS clustering algorithm, and visualizing the results using dimensionality reduction techniques, namely UMAP (Uniform Manifold Approximation and Projection) and t-SNE (t-Distributed Stochastic Neighbor Embedding). The primary objective was to assess the effectiveness of OPTICS in clustering high-dimensional cytometry data and to evaluate the resulting clusters in terms of clinical relevance. While the primary objectives were partially met, there were several challenges and limitations in the study. The OPTICS algorithm was able to identify distinct clusters within the data, as demonstrated by the reachability plots and scatter plots generated from the analysis. However, the clarity and distinctiveness of these clusters were not as strong as initially hoped. Many clusters appeared diffuse or overlapping, which hindered the ability to confidently interpret them as distinct immune cell populations. The study systematically explored different parameters for the OPTICS algorithm, including **xi, min_samples, max_eps, and min_cluster_size**. These parameters significantly impacted the clustering results, with some settings leading to over-segmentation or under-segmentation of the data. Despite extensive tuning, the resulting clusters were not always biologically meaningful, suggesting that further refinement or alternative approaches may be needed. Both UMAP and t-SNE were used to visualize the high-dimensional data in two dimensions. While UMAP provided a broader overview of the clustering structure, t-SNE excelled in revealing local relationships. However, the choice of distance metric (Manhattan vs Euclidean) also played a crucial role, with each metric offering different insights into the data's structure. The results varied significantly depending on the chosen parameters, which posed a challenge in selecting the best visualization for interpretation. Due to time constraints and the less-than-optimal quality of the clusters, differential analysis was not performed. Attempting to analyse poorly defined clusters would likely yield inconclusive or misleading results, and therefore, it was considered more sensible to focus on refining the clustering process before conducting further analysis.

6.1 Limitations

The use of the OPTICS algorithm demonstrated its flexibility in handling high-dimensional cytometry data without requiring a predefined number of clusters. The combination of UMAP and t-SNE provided complementary perspectives on the clustering results, allowing for a more comprehensive understanding of the data. However, the primary weakness was the difficulty in obtaining well-defined clusters that could be confidently interpreted as distinct immune cell populations. The sensitivity of OPTICS to parameter settings, coupled with the high-dimensional and noisy nature of the data, made it challenging to identify biologically meaningful clusters. Additionally, only one random seed was used for t-SNE, which could have influenced the stability and reproducibility of the visualizations. The Key lessons I learned, include the importance of extensive parameter tuning and the need to explore different distance metrics and how to visualize and analyse the resulted cluster. It was also noted that the choice of seed in t-SNE can significantly impact the results, suggesting that multiple seeds should be tested to ensure robustness.

6.2 Future Work

The Future work should revisit the differential analysis of the identified clusters once the clustering process is further refined. This analysis could provide valuable insights into the immune cell populations associated with RA flares and remission, potentially leading to the identification of novel biomarkers. Also, exploring other clustering algorithms or modifications to the OPTICS algorithm could improve cluster quality. For example, trying different distance functions or applying weights to specific biomarkers could enhance the separation of clusters. Experimenting with different data transformation techniques or scaling methods could help in better capturing the underlying structure of the data. This might involve using alternative transformations like logarithmic scaling or applying feature selection methods to reduce noise. In the end, the ultimate goal of this research is to contribute to the development of personalized treatment strategies for RA patients. By refining the clustering process and identifying clinically relevant biomarkers, future studies could pave the way for more targeted and effective therapies.

7 Conclusion

This dissertation explored the application of the OPTICS algorithm to flow cytometry data from RA patients, aiming to identify distinct immune cell populations and biomarkers relevant to disease activity. While the study successfully implemented OPTICS and used UMAP and t-SNE for visualization, the resulting clusters were not as distinct as expected, limiting the potential for robust differential analysis. This limitation highlights the challenges of applying density-based clustering to high-dimensional, complex biological data. The study emphasizes the need for further parameter tuning and possibly alternative clustering methods to improve cluster clarity and biological relevance. Despite these challenges, the research contributes valuable insights into the use of advanced computational techniques in immunology, particularly for RA. Future work should focus on addressing these limitations, including exploring different data preprocessing techniques, distance metrics, and clustering algorithms. Moreover, efforts should be made to complete the differential analysis and validate the clusters identified, ultimately aiming to advance personalized treatment strategies for RA patients.

References

1. Rayner, F., Anderson, A.E., Baker, K.F. et al. BIOlogical Factors that Limit sustAined Remission in rhEumatoid arthritis (the BIO-FLARE study): protocol for a non-randomised longitudinal cohort study. *BMC Rheumatology* **2021**, 5, 22. DOI: 10.1186/s41927-021-00194-3.
2. Edilova MI, Akram A, Abdul-Sater AA. Innate immunity drives pathogenesis of rheumatoid arthritis. *Biomed J.* 2021 Apr;44(2):172-182. DOI: 10.1016/j.bj.2020.06.010. Epub 2020 Jul 8. PMID: 32798211; PMCID: PMC8178572.
3. Scott, D. L., Wolfe, F., & Huizinga, T. W. J. (2010). Rheumatoid arthritis. *The Lancet*, 376(9746), 1094-1108. DOI: 10.1016/s0140-6736(10)60826-4.
4. Jang S, Kwon EJ, Lee JJ. Rheumatoid Arthritis: Pathogenic Roles of Diverse Immune Cells. *Int J Mol Sci.* 2022 Jan 14;23(2):905. DOI: 10.3390/ijms23020905. PMID: 35055087; PMCID: PMC8780115.
5. Scherer, H. U., Häupl, T., & Burmester, G. R. (2020). The etiology of rheumatoid arthritis. *Journal of Autoimmunity*, 110, 102400. DOI: 10.1016/j.jaut.2019.102400.
6. Zhao J, Guo S, Schrodi SJ, He D. Molecular and Cellular Heterogeneity in Rheumatoid Arthritis: Mechanisms and Clinical Implications. *Front Immunol.* 2021 Nov 25;12:790122. DOI: 10.3389/fimmu.2021.790122. PMID: 34899757; PMCID: PMC8660630.
7. Sriphum, W., Wills, G. B., & Green, N. G. (2022). Floptics: A Novel Automated Gating Technique for Flow Cytometry Data. *International Journal of Organizational and Collective Intelligence (IJOCI)*, *12*(1), 1-21. DOI: 10.4018/ijoci.301561.
8. Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *SIGMOD Record*, *28*(2), 49–60. DOI: 10.1145/304181.304187.
9. Roederer, M. (2008). How many events is enough? Are you positive? *Cytometry Part A*, *73A*(5), 384–385. DOI: 10.1002/cyto.a.20549.
10. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, *111*(26). DOI:10.1073/pnas.1408792111.
11. Rangaprakash D., Odemuyiwa T., Narayana Dutt D., Deshpande G.; Alzheimer's Disease Neuroimaging Initiative. (2020). Density-based clustering of static and dynamic functional MRI connectivity features obtained from subjects with cognitive impairment. *Brain Informatics*, *7*(1):19. DOI: 10.1186/s40708-020-00120-2. PMID: 33242116; PMCID: PMC7691406.
12. Aghaeepour, N., Finak, G., et al. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, *10*(3), 228-238. DOI: 10.1038/nmeth.2365.
13. McInnes, L., Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv (Cornell University)*. DOI: 10.48550/ARXIV.1802.03426.
14. van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.
15. Poon, M. M. L., Farber, D. L. (2020). The Whole Body as the System in Systems Immunology. *iScience*, *23*(9), 101509. DOI: 10.1016/j.isci.2020.101509. PMID: 32920485; PMCID: PMC7491152.
16. Firestein, G. (2003). Evolving concepts of rheumatoid arthritis. *Nature*, *423*, 356–361. DOI: 10.1038/nature01661.
17. Baker, K.F., McDonald, D., Hulme, G., et al. (2024). Single-cell insights into immune dysregulation in rheumatoid arthritis flare versus drug-free remission. *Nat Commun*, *15*, 1063. DOI: 10.1038/s41467-024-45213-2.
18. Jesus, J. (2022). Essential elements as critical players against SARS-CoV-2 activity. *Journal of Integrated OMICS*, *12*, 3-9. DOI: 10.5584/jiomics.v12i2.216.
19. Guo, Q., Wang, Y., Xu, D., et al. (2018). Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Res*, *6*, 15. DOI: 10.1038/s41413-018-0016-9.

20. Ding, Q., Hu, W., Wang, R., et al. (2023). Signaling pathways in rheumatoid arthritis: implications for targeted therapy. *Sig Transduct Target Ther*, *8*, 68. DOI: 10.1038/s41392-023-01331-9.
21. Zhao, J., Guo, S., Schrodi, S. J., & He, D. (2021). Molecular and Cellular Heterogeneity in Rheumatoid Arthritis: Mechanisms and Clinical Implications. *Front Immunol*, *12*, 790122. DOI: 10.3389/fimmu.2021.790122. PMID: 34899757; PMCID: PMC8660630.
22. Robinson, J. P., Ostafe, R., Iyengar, S. N., Rajwa, B., & Fischer, R. (2023). Flow Cytometry: The Next Revolution. *Cells*, *12*(14), 1875. DOI: 10.3390/cells12141875. PMID: 37508539; PMCID: PMC10378642.
23. Maecker, H. T., Trotter, J., et al. (2012). Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology*, *12*(3), 191-200. DOI: 10.1038/nri3158.
24. Perfetto, S. P., Chattopadhyay, P. K., & Roederer, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology*, *4*(8), 648-655. DOI: 10.1038/nri1416.
25. Saeys, Y., Van Gassen, S., & Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, *16*(7), 449-462. DOI: 10.1038/nri.2016.56.
26. Finak, G., et al. (2014). OpenCyto: An open-source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol*, *10*(8), e1003806. DOI: 10.1371/journal.pcbi.1003806.
27. Verschoor, C. P., Lelic, A., Bramson, J. L., Bowdish, D. M. E. (2015). An introduction to automated flow cytometry gating tools and their implementation. *Front. Immunol.*, *6*:380. DOI: 10.3389/fimmu.2015.003
28. Zindler, E., Schobel, A., et al. (2020). Immune cell subsets associated with systemic lupus erythematosus (SLE) in the adaptive immune system identified through clustering techniques. *PLoS ONE*, *15*(10), e0240731. DOI: 10.1371/journal.
29. De Jager, P. L., Rossin, E., Pyne, S., Tamayo, P., Ottoboni, L., Viglietta, V., Weiner, M., Soler, D., Izmailova, E., Faron-Yowe, L., O'Brien, C., Freeman, S., Granados, S., Parker, A., Roubenoff, R., Mesirov, J. P., Khoury, S. J., Hafler, D. A., Weiner, H. L. (2008). Cytometric profiling in multiple sclerosis uncovers patient population structure and a reduction of $CD8_{low}$ cells. *Brain*, *131*(7), 1701–1711. DOI: 10.1093/brain/awn118.
30. Couloume, L., Ferrant, J., Le Gallou, S., Mandon, M., Jean, R., Bescher, N., Zephir, H., Edan, G., Thouvenot, E., Ruet, A., Debouverie, M. (2021). Mass cytometry identifies expansion of T-bet$^+$ B cells and CD206$^+$ monocytes in early multiple sclerosis. *Frontiers in Immunology*, *12*, 65357.
31. Ankerst, M., Breunig, M. M., Kriegel, H.-P., Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data (SIGMOD '99)* (pp. 49–60). Association for Computing Machinery, New York, NY, USA. DOI: 10.1145/304182.304187
32. Redinger, G., Hunner, M. (2017). Visualization of the Optics Algorithm. In *VIS 2017 - Universität Wien.*
33. McInnes, L., & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (Cornell University). DOI: 10.48550/ARXIV.1802.03426.