atmosphere

MDPI

Article

Comparative Analysis of Traditional and Advanced Clustering Techniques in Bioaerosol Data: Evaluating the Efficacy of K-Means, HCA, and GenieClust with and without Autoencoder Integration

Maxamillian A. N. Moss [1], Dagen D. Hughes [2], Ian Crawford [1], Martin W. Gallagher [1], Michael J. Flynn [1] and David O. Topping [1,*]

[1] Department of Earth and Environmental Science, University of Manchester, Manchester M13 9PL, UK; maxamillian.moss@postgrad.manchester.ac.uk (M.A.N.M.)
[2] Droplet Measurement Technologies, Longmont, CO 80503, USA; dhughes@dropletmeasurement.com
* Correspondence: david.topping@manchester.ac.uk

Abstract: In a comparative study contrasting new and traditional clustering techniques, the capabilities of K-means, the hierarchal clustering algorithm (HCA), and GenieClust were examined. Both K-means and HCA demonstrated strong consistency in cluster profiles and sizes, emphasizing their effectiveness in differentiating particle types and confirming that the fundamental patterns within the data were captured reliably. An added dimension to the study was the integration of an autoencoder (AE). When coupled with K-means, the AE enhanced outlier detection, particularly in identifying compositional loadings of each cluster. Conversely, whilst the AE's application to all methods revealed a potential for noise reduction by removing infrequent, larger particles, in the case of HCA, this information distortion during the encoding process may have affected the clustering outcomes by reducing the number of observably distinct clusters. The findings from this study indicate that GenieClust, when applied both with and without an AE, was effective in delineating a notable number of distinct clusters. Furthermore, each cluster's compositional loadings exhibited greater internal variability, distinguishing up to 3× more particle types per cluster compared to traditional means, and thus underscoring the algorithms' ability to differentiate subtle data patterns. The work here postulates that the application of GenieClust both with and without an AE may provide important information through initial outlier detection and enriched speciation with an AE applied, evidenced by a greater number of distinct clusters within the main body of the data.

Keywords: PBAP; bioaerosol; UVLIF; WIBS; machine learning (ML); cluster analysis; GenieClust; K-means; HCA; real-time detection and analysis; fungal spores; bacteria; pollen; climate change

Citation: Moss, M.A.N.; Hughes, D.D.; Crawford, I.; Gallagher, M.W.; Flynn, M.J.; Topping, D.O. Comparative Analysis of Traditional and Advanced Clustering Techniques in Bioaerosol Data: Evaluating the Efficacy of K-Means, HCA, and GenieClust with and without Autoencoder Integration. *Atmosphere* **2023**, *14*, 1416. https://doi.org/10.3390/atmos14091416

Academic Editors: David J. O'Connor, Eoin McGillicuddy, Meheal Fennelly and John R. Sodeau

Received: 9 August 2023
Revised: 30 August 2023
Accepted: 5 September 2023
Published: 8 September 2023

1. Introduction

Bioaerosols are airborne particles comprising living organisms or their components. Varying in both size and composition, they can include a variety of microorganisms such as fungal spores (2–50 µm) [1], bacteria (0.2–10 µm) [2], pollens (10–150 µm) [3], and viruses (0.022–0.15 µm) [4]. Bioaerosols play a significant and multifaceted role in modern society having both positive and negative impacts on human life and the environment. These can range from smaller scale topics such as occupational health [5] and indoor air quality [6] to larger scale themes such as allergies [7], agricultural productivity [8], and environmental air quality [9]. Understanding the roles of bioaerosols in society has thus been recognised as fundamental for safeguarding public health, food resources, and the environment in a landscape that is continuously facing increasing pressures from population growth, urban development, and climate change.

Historically, the study of bioaerosols has relied on traditional 'offline' methods involving a slew of laboratory-based approaches involving the culturing of bacteria for genome

sequencing [10] or microarrays [11], and visually examining morphological features of pollens or fungal spores using microscopy [12]. However, the late 20th century brought with it technological developments that facilitated the use of 'real-time' ultra violet light-induced fluorescence (UVLIF) instruments capable of recording a particle's fluorescence, size, shape, and asymmetry in situ [13].

Whilst achieving precise speciation remains a persistent challenge [8], which some recent studies have sought to address [14], the introduction of UVLIF instruments nonetheless represented a significant advancement in the field, and these continue to be developed to this day. These instruments have enabled the collection of extensive datasets spanning vast geographical distances and time periods [13,15–17]. The Wideband Integrated Bioaerosol Sensor (WIBS) stands out as a prominent and widely recognised UVLIF instrument that has been extensively utilized in various scientific investigations over the years with the WIBS-NEO/5 representing their current flagship model for bioaerosol detection. For instance, from an early 2010 field study, where the WIBS unit was employed to measure Primary Biological Aerosol Particles (PBAP) in a tropical canopy environment [18], to one of its most recent applications during an 18-month environmental campaign conducted in Sweden [19]. Subsequent to the advancement in UVLIF, the datasets collected are often utilized in conjunction with conventional clustering algorithms such as K-means or hierarchal clustering to construct temporal profiles characterized by distinct clusters corresponding to different particle types [16].

However, despite these advancements, traditional clustering methods are increasingly proving to be restrictive in 'online' analysis. Specifically, in the case of the widely used and straightforward algorithm, K-means, limitations arise from its assumption of spherical clusters, its inability to handle outliers, and its failure to capture complex relationships and nonlinear structures, among other factors [20,21]. Similarly, hierarchal cluster analysis (HCA), whilst offering more intricate analyses, presents limitations in terms of its sensitivity to noise and the high computational cost it entails [22,23].

The recent surge in atmospheric science research has unveiled a myriad of machine learning (ML) approaches [24–26]. For instance, artificial neural networks (ANN) and recurrent neural networks (RNN) have been applied with marked success in domains such as wastewater treatment projections and haze prediction models [27,28]. Chen et al. (2023) utilized the back propagation (BP) neural network to assess the impact of meteorological conditions on haze, underscoring its high accuracy, minimal errors, and efficiency even with limited learning data [26]. However, this method, as with others, necessitates training data, which limits its real-time application in ambient studies where, most often, one does not have good label data or predefined categories. This issue can be further exacerbated if, such was the case with one recent study utilizing convolutional neural networks (CNN), there is an insufficient amount of training information. Notwithstanding their merits, direct use of supervised methods restricts their application in clustering bioaerosols and is wholly reliant on robust, and often extensive, training libraries. Despite these advancements to supervised approaches, there remains a lack of research dedicated to the development of unsupervised algorithms for clustering of ambient bioaerosols. Consequently, the more antiqued algorithms, including K-means and HCA, continue to dominate the field [29–31].

Amidst this backdrop, GenieClust has emerged as a potentially transformative, unsupervised algorithm for bioaerosol studies. This relatively novel and underexplored algorithm presents potential advantages due to its specialized focus on outlier detection and improved computational performance [32]. Similarly, the use of autoencoders (AE), an emerging deep learning model, holds promise as it can effectively encode substantial quantities of data into a latent vector, enabling the extraction of meaningful features whilst potentially reducing the computational costs involved and improving the performance of simpler methods such as K-means [33].

In this study, we present observations of fluorescent particle concentrations and size obtained during a field campaign for the Observation System for Clean Air (OSCA) initiative in central Manchester, using a continuous multiwavelength UV-LIF aerosol spectrometer.

Our analysis employed several conventional and emerging machine learning techniques to differentiate between various clusters of observed particles, and user expert interpretation to examine the differences and similarities between the techniques themselves. As far as we can infer, this is the first study to combine traditional and emerging deep learning and clustering methods for this application. To gain an insight into the type of particles witnessed during the study, a supplementary single particle analysis was also carried out utilising a well-established 'ABC' classification system [34]. Additionally, we used meteorological data obtained from the same campaign to examine any relationships between the events identified and the concurring environmental conditions. Furthermore, we explore how using the particle-by-particle output from the instrument's quadrant detector may influence the clustering results, which until now, has only been used to yield a single asymmetry value (see Section 2.2).

In this paper, we provide a case for new and emerging ML techniques as a suitable replacement to conventional methods, potentially offering a greater degree of accuracy and flexibility whilst also reducing the time-to-solution for unsupervised classification. We caveat this statement by the need for expert interpretation of extracted clusters, including inferring the optimum clusters available. Whilst methods for determining the optimum number of clusters have been used in previous studies [35–37], our results show that there is a danger of missing increasingly diverse information in higher numbers of clusters that would otherwise be aggregated using automated methods.

2. Materials and Method

Figure 1 provides a summary of the steps involved in this study, from acquiring data to generating cluster solutions. The corresponding section additionally outlines each step below for reference. The following section outlines the preprocessing and cleaning strategies, in addition to data normalisation, and clustering techniques.

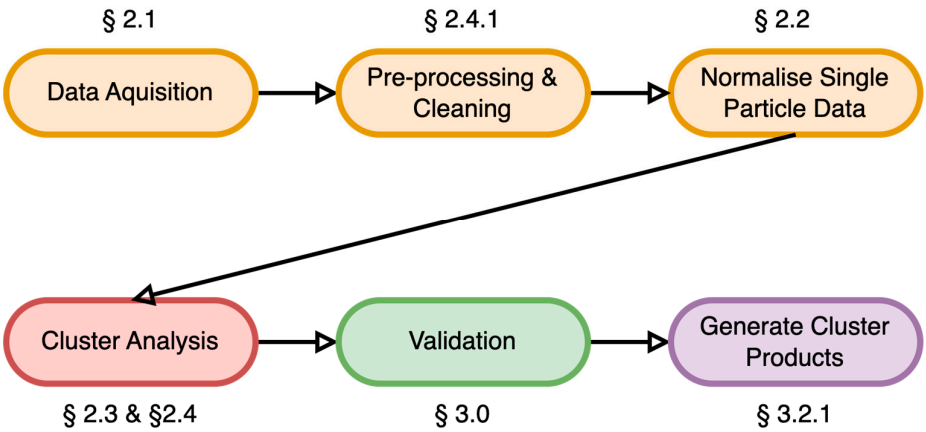

Figure 1. Flowchart illustrating the steps to generate cluster products from single particle data. §: Refers to the corresponding section of the paper. Adapted from Crawford et al. (2015) [23].

2.1. Sampling

Bioaerosol measurements spanning the duration of the experiment (6 June to 14 July 2021) were acquired at the Manchester Air Quality Supersite situated at the Firs Environmental Research Station (53°27′ N, 2°13′ W) in Manchester, United Kingdom (http://www.cas.manchester.ac.uk/restools/firs/, accessed on 17 May 2023). The sampling site itself is positioned approximately 4 km from the city centre. Manchester, a thriving urban region, is located in the northwest of England and is the second most densely populated area within the United Kingdom. Nevertheless, the city encompasses numerous verdant areas. Size and fluorescence measurements were obtained using the latest Wideband Integrated Bioaerosol Sensor (WIBS, 0.5–30 µm particle range, model: 5/NEO, Droplet Measurement Technologies LLC., 0.3 L min^{-1} sample flow and 2.1 L min^{-1} sheath flow).

The device employs a 635 nm laser to gauge optical particle dimensions and asymmetry. Upon activation by the sizing laser signal, two xenon flashlamps emit light that is filtered at 280 nm and 370 nm to simulate fluorescence. Subsequently, fluorescence is captured across three bands using two broad-range detection channels (310–400 nm and 420–650 nm). A summary of the instrument's wavelengths and bands can be found in Table 1.

Table 1. The excitation wavelengths and detection wavebands for each of the three fluorescent (FL) channels of the WIBS-NEO/5.

	FL1 280		**FL2 280**		**FL2 370**	
	Excitation	**Detection**	**Excitation**	**Detection**	**Excitation**	**Detection**
WIBS-NEO/5	280 nm	310–400 nm	280 nm	420–650 nm	370 nm	420–650 nm

Temperature, rainfall, and humidity were measured at the supersite meteorological station using a solid-state probe in an aspirated radiation shield (recorded using a Palas FIDAS 200), laser disdrometer, and aspirated capacitance probe, respectively [38].

2.2. Preprocessing

Original data files from the WIBS-NEO are distributed in Hierarchical Data Format 5 (HDF5) format. The recorded data within each file comprises fluorescence intensity (refer to Table 1), particle size, asymmetry factor (AF), and monitoring data. The AF is determined through four measurements obtained using a quadrant detector and inserted into a derived Mie-theory calculation to yield the AF value. To facilitate analysis, the WIBS-NEO employs Unix time as the timestamp for its data, which is later converted to date/time format. Prior to the commencement of data collection, a Forced Trigger (FT) action is performed, wherein the instrument carries out preliminary data collection within an empty chamber without any throughflow. This step establishes baselines that serve as the fluorescent threshold for subsequent analysis [35].

Previous studies have iterated on preprocessing steps that remove particles with a fluorescence intensity outside of a set number of standard deviations from the mean [16,35,39]. This has been used previously to remove the impacts of interferants, weakly fluorescing particles, and thus, potentially increase the sensitivity of clustering methods. Thus, this approach is adopted here with the threshold being set using an averaged FT and three times the standard deviation (3 σ).

Prior to any clustering or deep learning application, each variable was normalised using a combination of z-score, log transforms, and power transforms to generate normal distributions around a mean of 0 as standard practice for distance-based methods [35].

2.3. Analytical Techniques

2.3.1. Cluster Analysis

Machine learning offers a route for converting an instrument response into distinct groups or, more specifically, bioaerosol particles measured. The ideal outcome would be labelling each observed particle into higher level categories including pollen, fungal spores, bacteria, and interferents such as dust or heavy metals [34,40]. Supervised learning algorithms are trained to data obtained from controlled studies in which particle types are known a priori. For example, this could include resuspension of multiple pollen types within an environment chamber. There are multiple supervised learning algorithms available, from boosted forest methods such as XGBoost [41] through to more complex deep learning approaches [42]. Unfortunately, previous work highlighted difficulties in sample generation, particularly with regards to fungal spores and bacteria [43], leading to potential inaccuracies in predicted labels. Whilst these issues remain, and given the complexity of the source–receptor relationships in ambient bioaerosols, we instead evaluate the utility of unsupervised methods in this study.

Unsupervised learning algorithms enable us to convert the measured instrument response into distinct groups, or 'clusters'. Cluster analysis has been used in many previous UVLIF studies to varying levels of success [23,35,37,44,45]. As with supervised approaches, there are a range of algorithms available. Previous work using the WIBS within controlled setting has, despite issues around sample generation, led to the adoption of HCA for use in ambient studies. Hierarchal cluster analysis is a distance-based method in which a pair of observations are combined according to the lowest 'distance' between this pair, based on a matrix of distances between all observations. HCA then sequentially adds additional observations based on the lowest distance and starts to form groups, or clusters. If an ambient dataset holds 10,000 observations, HCA begins with each observation as a separate cluster and sequentially merges towards one cluster. Whilst there are metrics to better understand the optimum number of clusters, the labelling of each cluster to a specific bioaerosol type requires expert interpretation using information related to metrics such as diurnal cycles, particle data [46], compositional loadings [34], and response to humidity and rainfall [23].

In this study, we used a number of clustering algorithms based on two overarching challenges when applied to bioaerosol datasets:

Data quantity: Distance-based learning algorithms can become infeasible when the data quantity is large (see Section 2.4.3). It is not uncommon for ambient datasets to contain millions of observations. Generating pairwise distances for large datasets can reach limits of available memory. Whilst methods have been developed to circumvent this, and used in our study and detailed shortly, the time-to-solution can still remain prohibitive, with orders of weeks required to arrive at a clustering solution for any given preprocessing strategy.

Optimum cluster selection: Whilst metrics exist to determine the optimum number of clusters from a clustering algorithm, there are pros and cons of each method. These not only relate to the characteristics of the data, and thus preprocessing, but also data quantity. Indeed, metrics can also present a prohibitive time-to-solution, which places further dependencies on the time taken by the underlying clustering approach.

Given the above, in this study, we evaluate several emerging methods and strategies designed to not only lead to an improved time-to-solution, but potentially increased insights into the particles sampled. In the following section we detail the various approaches taken, which are summarised in Table 2.

Table 2. A summary of the analytical approaches taken within this study. Level 1 denotes whether quadrant information from the WIBS instrument was used; Level 2 outlines the data standardization method used; Level 3 describes the cluster method undertaken.

Level 1. Data Selection									
-	No Quad	**x**	**x**	**x**	**x**	**x**	**x**		
-	Quad							**x**	**x**
Level 2. Data Standardisation									
-	Robust Scaler/Log transform	**x**		**x**		**x**		**x**	
-	Deep Autoencoder		**x**		**x**		**x**		**x**
Level 3. Cluster Technique									
-	HCA	**x**	**x**						
-	K-means			**x**	**x**				
-	GenieClust					**x**	**x**	**x**	**x**

The methods are built around a number of data and algorithmic procedures. At the first level, we distinguish between outcomes using the quad channel's signal intensity or just the AF, which, as described in Section 2.2, is a metric to understand the asymmetry of the detected particle. At the second level, we can use the standardised data generated from transformations described in Section 2.4.1. We refer to this as 'raw' in the context of retaining the total number of metrics used. At this same level, we can also choose to compress the original number of metrics, thus dimensions of our data, before passing through any of the clustering algorithms. Dimension reduction techniques vary from linear to nonlinear [47]. In this study, we chose a nonlinear approach by fitting and deploying deep learning autoencoders to our dataset. Whilst the WIBS does not generate many dimensions, AEs have been shown to improve the performance of clustering methods, including K-means, in a number of applications [48–50]. Following this, at the third level, we employ a range of clustering approaches and evaluate each through a side-by-side comparison of cluster properties.

2.3.2. Optimal Cluster Selection

Clusters have herein been determined via expert manual interpretation. This is advantageous when compared to the application of metrics such as the silhouette score for several reasons. First and foremost, it allows for the integration of domain knowledge and expertise, which may not be considered through the use of metrics exclusively. Furthermore, where most clustering metrics assume a certain structure to the data, which as this study will show can pose complications, human interpretation does not. However, many more advantages are evident, such as providing contextual relevance to the results as well as applying human judgement and intuition, which, by the very nature of unsupervised clustering, can be beneficial.

The optimal number of clusters is determined by the criterion of reaching a point where no new clusters are identified and overlapping clusters start to emerge. This ensures that the clustering process has captured all meaningful variations in the data.

The distinct number of clusters here refers to the number of unique clusters within the optimal cluster number. For example, an optimal number of clusters equalling eleven and a distinct number of clusters equalling eight would mean there were three overlapping clusters identified that were considered to be the same as another cluster(s). An equation to represent this example is shown below:

$$C_{Overlapping} = C_{Optimal} - C_{Distinct}$$

Differentiating between overlapping clusters and distinct clusters is critical for the interpretation stage of the analysis. Relying on the provided example, a lack of such differentiation could lead to the misrepresentation that the algorithm has pinpointed eleven separate clusters. This could misrepresent both the algorithms' capabilities and the actual bioaerosol types and subcategories under investigation. The selection of a distinct number of clusters is guided by several methodical criteria. Firstly, differences in the temporal profiles are considered, including the number of particles and the occurrences of spikes at specific times of the day. Secondly, variations in size and the interquartile range are considered. This helps identify clusters that differ in physical dimensions. Moreover, differences in compositional loading, which refers to the distribution of specific fluorescent characteristics or attributes within a cluster, are considered. Lastly, clusters exhibiting similarities in size, diurnal profile, and compositional loadings are categorised as representing the same type of particle, as these shared characteristics indicate a common underlying nature. By considering these multiple criteria, a more comprehensive understanding of the clusters and their similarities/differences can be achieved. A summary of the optimal and distinct number of clusters for each evaluated method is displayed in Table 3.

Table 3. The optimal number of clusters and the number of clearly defined clusters, distinct, within that optimal number for each of the examined clustering methods. Raw refers to the clustering methods being applied directly to the single particle data; AE signifies the methods being applied to the autoencoded data (the 'latent space', or 'latent').

	K-Means		HCA		GenieClust	
	Raw	**AE**	**Raw**	**AE**	**Raw**	**AE**
Optimal No. Clusters	6	6	8	6	8	11
Distinct Clusters	6	6	7	6	8	8

2.4. Methods Evaluated

Each clustering method requires a preprocessing stage, where the data are standardised to remove the impacts of variable scales for each metric. Distance-based clustering provides several strategies for combining sequential clusters. For the optimised variant of HCA used in this study, we use the 'single linkage' strategy to benefit from the memory efficient version given the number of observations in this study.

For both the transformed (or 'raw') and autoencoded GenieClust approaches, we generate 2 to 20 clusters. To assess the accuracy of metrics on this new method, the Calinski–Harabasz score is used to infer the optimum number of clusters [23], supplementary to the manual interpretation.

In this study, the terms 'transformed' and 'raw', in addition to 'autoencoded (AE)' and 'latent', are used synonymously.

2.4.1. Preprocessing Strategies

Scaling data before any clustering algorithm is important. We use two general approaches. In the first, we transform each original data product captured by the instrument. More specifically, for the fluorescent intensity, a log transform is applied before using the RobustScaler transformation provided by the Scikit-learn package [51,52]. This approach removes the median and scales the data according to the quantile range.

As a second option, we also use AEs to compress the original number of dimensions prior to clustering. Whilst we do not have many dimensions, previous research has demonstrated their ability to improve outcomes from clustering methods. AEs are a family of methods emerging from deep learning neural networks. The basic premise is the neural network takes an input, and compresses and then expands these data through a series of layers, with the aim that the output is as close to the original input as possible (Figure A1). The smallest layer, in our case positioned in the middle of this network, is a 'latent' or 'encoded' representation of our data. For example, in our case, we could have 1 dimensional array of 9 entries that represent the 3 fluorescent channels, asymmetry factor, size, and four quad channel intensities. The first layer of the AE would have a smaller number of nodes and the latent layer perhaps 3 nodes. There are a number of hyperparameters to tune, from the number of layers, drop-out strategy, and choice of activation function in each node. For our work, we use the Keras Hypertune package in Python [51] to determine the optimum architecture of the AE, using a mean squared loss between the input and output. Once optimised, the latent space is then used as input into subsequent clustering methods. In the following section, we discuss the clustering techniques.

2.4.2. K-Means

K-means is a widely used method for cluster analysis and is available in a number of platforms. Performance metrics highlight how much quicker K-means is than other clustering methods, offering an ability to work with very large datasets [53]. However, K-means is also limited in its ability to work with data with complex characteristics. This can be seen even for relatively simple 2D datasets that can be separated by visual inspection [52]. K-means does not produce an optimum number of clusters and requires the use of cluster validation indices.

2.4.3. HCA

Hierarchal cluster analysis is a widely used clustering method that allows for the identification of clusters within a dataset based on the similarity or dissimilarity of its objects. This is achieved by progressively merging or dividing clusters in a hierarchal structure until a desired level of granularity is achieved [54]. However, HCA has certain limitations. One major limitation is its sensitivity to noise. HCA is often influenced by outliers or noisy data points, which can result in the formation of suboptimal or incorrect clusters. Another limitation is the high computational cost associated with HCA, especially when dealing with large datasets such as the one in this study. The hierarchal nature of the algorithm requires pairwise distance calculations between all objects, making it computationally intensive and time-consuming.

2.4.4. GenieClust

'GenieClust' is a relatively new clustering algorithm that, to the best of our knowledge, remains largely unused in atmospheric studies. According to the documentation, GenieClust performs hierarchical clustering of datasets with millions of points 'within minutes', provided the data can fit within the memory [55]. Comparisons with the 'fastcluster' package, also used in this study, suggest time-to-solution improvements. GenieClust presents both an 'accurate' and an 'approximate' method that relate to strategies used for combining clusters, with the former method offering more performance enhancement. We use the exact method in this study to offer a fair comparison between all clustering methods used.

2.4.5. ABC

The WIBS-NEO utilizes 2 excitation wavelengths and 3 detection wavebands, summarized in Table 1. These are broadly categorized as fluorescence channels 1 (or A), 2 (or B), and 3 (or C). Different types of particles have been shown to generally exhibit dissimilar fluorescent characteristics that when using the 'ABC' method in conjunction with other metrics, can infer the type or group of particles being observed. As aforementioned, the 3 channels represent fluorescence in a single category, A, B, or C. Fluorescence in two channels is therefore designated as either AB, AC, or BC, whilst fluorescence in all 3 channels is labelled as ABC.

3. Results

3.1. K-Means and HCA

K-means, despite its simplicity, demonstrated exceptional performance compared to HCA in this study. Whilst HCA identified seven clusters (Figure 2), K-means revealed six distinct clusters in the transformed dataset (Figure 3). A comparative analysis of the temporal profiles of the transformed clusters generated by both K-means and HCA showed excellent agreement, with all six clusters matching in terms of profiles and sizes as identified by K-means. Interestingly, HCA identified a seventh, unique cluster with low concentrations averaging between 0.001 and 0.002 cm^3 and an average size of 0.3 µm (cluster 0; Figure 2). These characteristics suggest the possibility of it being an outlier within the dataset. This finding aligns with the existing literature, as K-means is known to struggle in accurately identifying outliers. The presence of this outlier cluster reinforces the notion that K-means may have limitations in effectively handling data points that deviate significantly from the majority of the dataset.

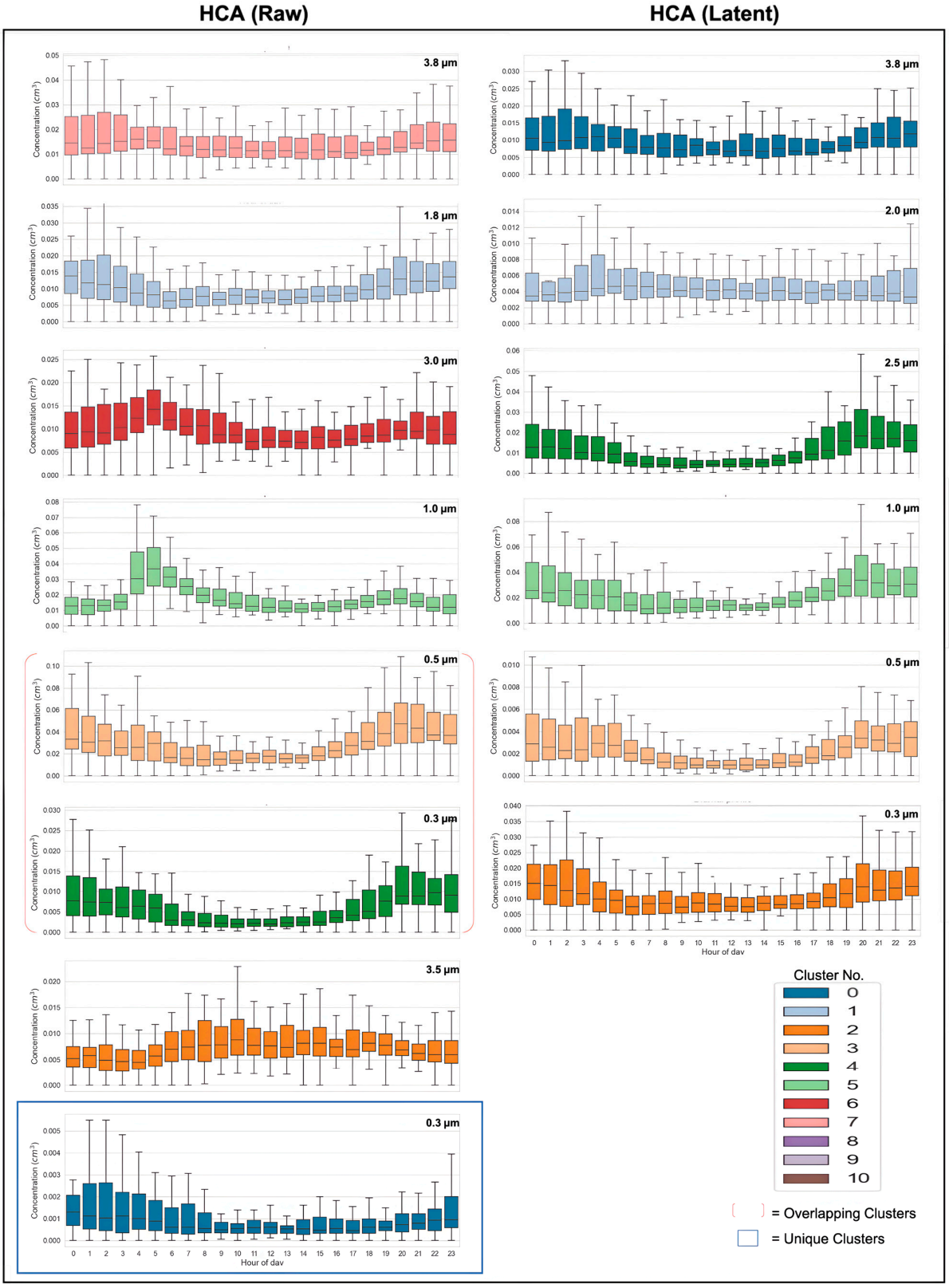

Figure 2. Temporal profiles exhibiting the average recorded particle concentrations (cm^3) for each hour of the day for both HCA approaches. The left column corresponds to HCA (raw), whilst the right column denotes HCA (latent). The average size of each cluster is presented in microns on each profile. The number of clusters for each method correspond to the optimal number of clusters identified by that method. Overlapping clusters are grouped with red parentheses. Unique clusters, which represent clusters that were only identified by the respective method, are grouped into blue boxes.

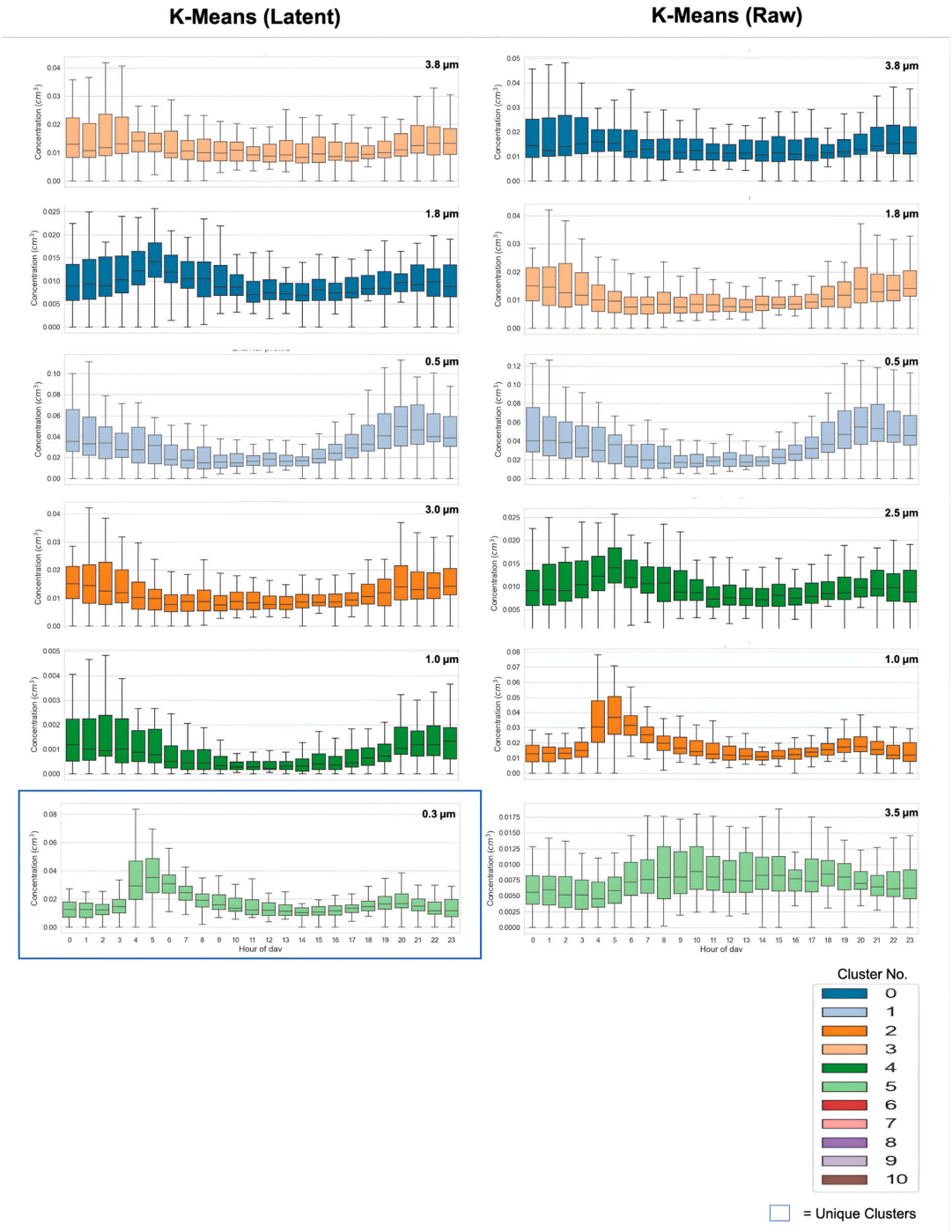

Figure 3. Temporal profiles exhibiting the average recorded particle concentrations (cm^3) for each hour of the day for both K-means approaches. The left column corresponds to K-means (latent), whilst the right column denotes K-means (raw). The average size of each cluster is presented in microns on each profile. The number of clusters for each method correspond to the optimal number of clusters identified by that method. Unique clusters, which represent clusters that were only identified by the respective method, are grouped into blue boxes.

The application of autoencoded data to the two methods resulted in notably distinct outcomes upon comparison. Specifically, when using K-means, the results appeared to align closely with the transformed approach, revealing the presence of six distinct clusters. Among these clusters, five exhibited similarities to the transformed clusters in terms of their size, concentrations, and diurnal profiles (see Figure 3). It is worth highlighting the

sixth cluster (cluster 5 (light green); Figure 3), which exhibited discrepancies between the two approaches. In the transformed cluster analysis, this cluster consisted of particles with an average size of approximately 3.5 µm. Notably, these particles displayed heightened activity levels from 8:00 am to 17:00 pm, with low concentrations ranging from 0.0025 to 0.0125 cm^3. Conversely, in the autoencoded cluster analysis, a different diurnal profile emerged, with activity peaking from 4:00 am to 7:00 am, and again from 19:00 pm to 20:00 pm. Furthermore, the average size of the particles in this cluster was approximately 0.3 µm, with concentrations ranging from 0.01 to 0.05 cm^3.

Intriguingly, these results exhibit similarities with a cluster, designated cluster 2, that emerged from the latent HCA (Figure 2). This cluster exhibits comparable characteristics in terms of composition (refer to Figure A2) and size (0.3 µm). However, distinctions in the average concentration and diurnal profile introduce some uncertainty to our understanding. Moreover, we observe a noteworthy aspect in the compositional loadings themselves. As illustrated in Figure A2, both of the 'raw' approaches primarily reveal clusters comprising a single particle type. This observation leads us to consider two possibilities: either the clusters indeed consist of only one particle type, or (and), more likely, the methods employed might have limitations in identifying clusters with more intricate compositions. Supporting the latter possibility, the latent outputs were analysed for both techniques, and they successfully differentiated three clusters, each containing multiple particle types.

Unlike K-means, the application of HCA on the autoencoded dataset appears to have exhibited no benefit in performance. Similar to the comparison conducted with K-means, HCA demonstrated good agreement with the transformed analysis, identifying five clusters that aligned well with previously identified clusters. However, a notable disparity emerged when comparing the autoencoded analysis to the transformed analysis. Firstly, the aforementioned outlier cluster, which showcased distinct characteristics in the transformed analysis, was not detected by HCA when applied to the autoencoded dataset, signifying the elimination or obscurement of distinguishing features during the autoencoding process. Secondly, the cluster with an average size of 3.5 microns and a peak activity throughout the day, which was evident in the transformed analysis for both algorithms, was also absent in the autoencoded analysis, further implying a loss or distortion of information. As a result, the AE HCA was left with six distinct clusters, which were identically matched to the six produced by AE K-means.

3.2. Genie Performance

The comparative analysis of GenieClust, K-means, and HCA on both datasets yielded compelling evidence of GenieClust's outlier detection capabilities, potentially surpassing those of traditional methods. Notably, GenieClust demonstrated a considerably higher number of optimal and distinct clusters compared to the conventional techniques (refer to Table 3). This enhanced performance is indicative of GenieClust's ability to discern finer patterns in the data, leading to more accurate and refined clustering results. Furthermore, through an ABC analysis of the clusters, we discovered a higher level of complexity in the compositional loadings of each cluster when utilizing GenieClust (see Figure A3). Intriguingly, each cluster identified by GenieClust exhibited at least three distinct particle types, demonstrating its capability to discriminate between clusters with a higher degree of precision compared to traditional methods. This aspect of GenieClust's performance is potentially significant, as it reveals how particle associations are multifaceted and how conventional methods may be unable to capture their complexity to the same degree. The implications of these findings are twofold. Firstly, comparing clusters identified by traditional and emerging methods becomes a more challenging task due to the superior discriminative abilities of GenieClust. For instance, we observed cluster 7 in both transformed analyses (Figures 2 and 4), where both GenieClust and HCA identified similarities in size (3.8 µm), diurnal profile, and to a small extent, composition. However, the distinct difference arises in the composition identification: whilst HCA exclusively categorized it as type B, GenieClust identified approximately 70% of the cluster as type B, along with smaller amounts of type A, C, and AB. This highlights the

subtleties that GenieClust is capable of capturing, which might be overlooked by traditional methods. Secondly, the greater ability of GenieClust to distinguish complex clusters indicates that the clusters identified by this algorithm are inherently slightly different from those identified by simpler, traditional algorithms. This implies that GenieClust offers a deeper and more nuanced understanding of the underlying patterns within the data, which can lead to more insightful interpretations and informed decision making in various scientific domains.

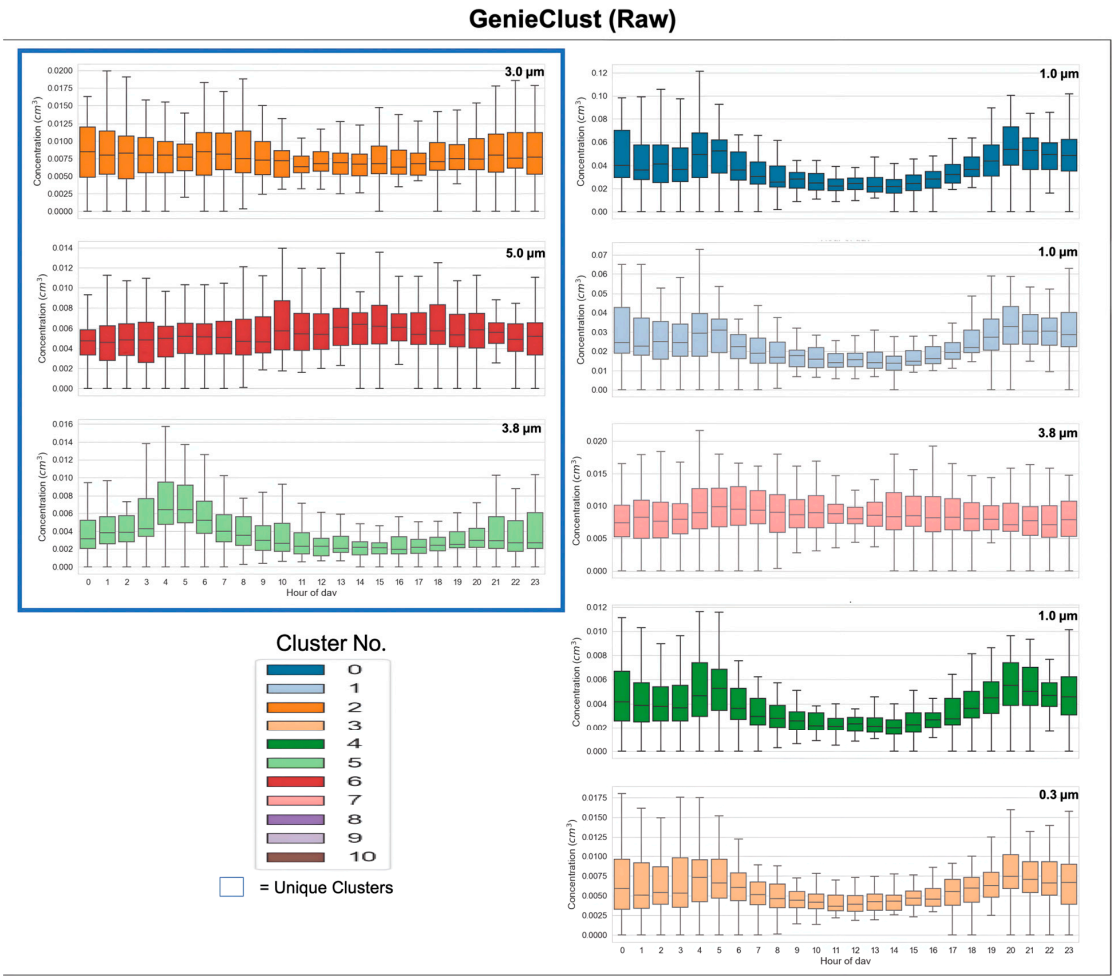

Figure 4. GenieClust (Raw) temporal profiles exhibiting the average recorded particle concentrations (cm^3) for each hour of the day. The average size of each cluster is presented in microns on each profile. Number of clusters correspond to the optimal number of clusters identified. Unique clusters, which represent clusters that were only identified by the respective method, are grouped into blue boxes.

The utilization of the autoencoder in conjunction with GenieClust has returned some unexpected differences to the transformed approach (Figure 5). Notably, the autoencoded GenieClust approach provided the highest number of optimal clusters but yielded the same number of distinct clusters to the transformed analysis (Table 3). Similar to the application of K-means, the AE analysis successfully identified a greater number of distinct clusters in the smaller size ranges (3 clusters), compared to the raw analysis. Additionally, the AE did not identify the three clusters from the transformed analysis, which exhibited larger sizes of 3 µm, 3.8 µm, and 5 µm. Just as described in the K-means breakdown (see Section 3.1), the missed clusters maintained consistent activity throughout the day and had small concentrations ranging from 0.002 to 0.0125 cm^3. This observation further solidifies the notion that the AE effectively filters out larger and less frequent particle data. However, this outcome can be viewed as advantageous since the AE analysis did resolve three additional distinct clusters within the lower size ranges (0.6 µm, 0.7 µm, and

2.5 µm). Furthermore, in addition to these differences, the two methods demonstrated good agreement on five other clusters.

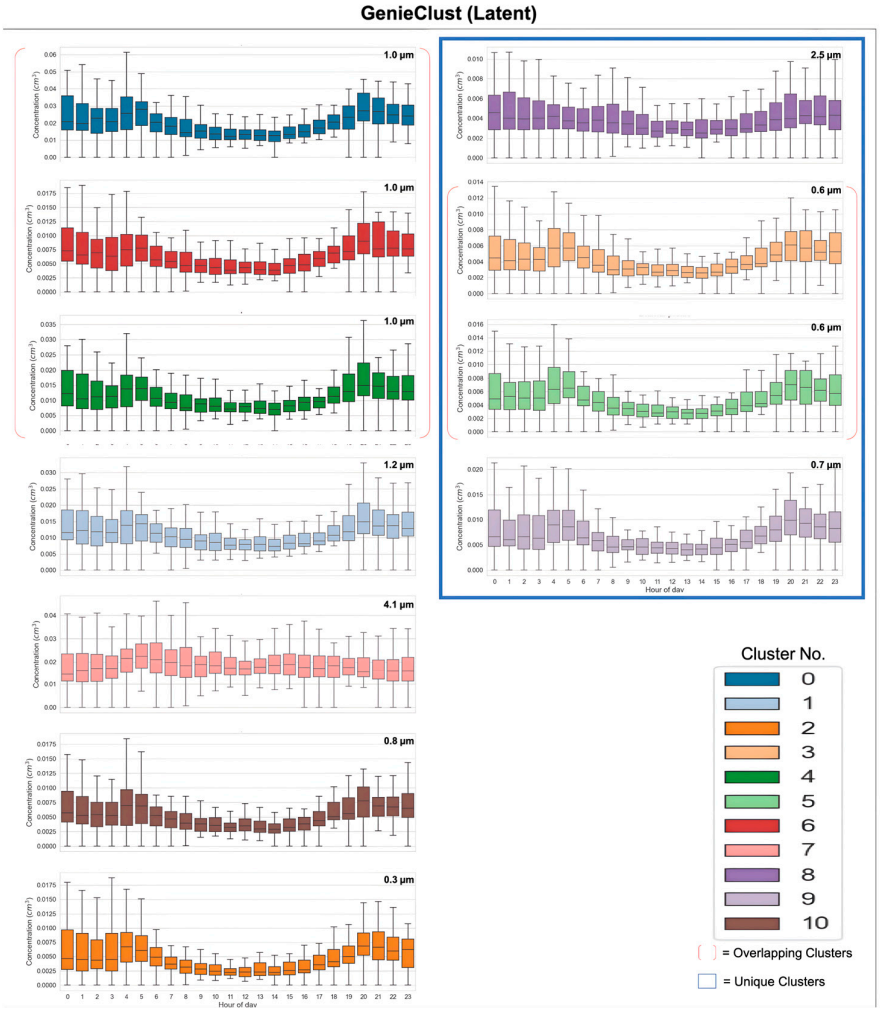

Figure 5. GenieClust (latent) temporal profiles exhibiting the average recorded particle concentrations (cm^3) for each hour of the day. The average size of each cluster is presented in microns on each profile. Number of clusters correspond to the optimal number of clusters identified. Overlapping clusters are grouped with red parentheses. Unique clusters, which represent clusters that were only identified by the respective method, are grouped into blue boxes.

Calinski–Harabasz Score

In addition to the manual interpretation, a Calinski–Harabasz score (CH) was performed comparing the two GenieClust approaches. Furthermore, to explore whether the addition of quadrant detector measurements (information which up until now has only been used to yield a single AF value) potentially improves clustering, the same algorithms were performed including this information and compared.

The relationship between the Calinski–Harabasz score value and number of clusters is illustrated in Figure 6. As can be observed from Figure 6A, the raw clustered data peak at three clusters before sharply dropping by several orders of magnitude to a potential peak between cluster 8 and 9. Conversely, the latent dataset appears to show prominent peaks at clusters 3, 6, and 11 before plateauing with a maintained high score, relative to the raw value. Figure 6B displays three peaks for both the raw and latent data at clusters 2, 5, and 8.5/9 and 3, 7.5, and 11, respectively. Upon assessing the CH-Score, there is an evident increase in the quality of the clusters when encoding the data to a latent vector, as is demonstrated by the consistently higher scores exhibited in Figure 6A,B. Observably, the raw and latent data

share an inverse relationship to one another for reasons not entirely clear. However, it should be noted that in almost all cases, the troughs displayed by the latent data retain a higher overall CH-Score than the peaks for the raw data, thereby inferring a greater degree of cluster separation regardless. Interestingly, despite the large decrease in the CH-Score in Figure 6B, a degree of similarity can be drawn from the optimal clusters observed for both latent outputs with clusters 3 and 11 being identified. From this we can surmise that the variance between outputs whether the quadrant information (QI) is included or not is relatively small and that the inclusion of it does not appear to provide better cluster separation. On the contrary, the inclusion of this information may contribute unnecessary 'noise'. This can be further corroborated upon visualizing a time series of the methods side by side (Figure A4). Indeed, observing 11 clusters over the duration of the study for both latent outputs, it is clear that a higher degree of similarity between clusters is present when including the QI (Figure A4B) as opposed to without (Figure A4A), suggesting several overlapping clusters. In particular, clusters 0, 1, 3, 4, 5, 6, 9, and 10 (eight in total) are all visually similar, in contrast to the raw latent output, which exhibits half the number of similarities (clusters 3, 4, 5, and 6). The sizes of each cluster, as shown in Figure A5, demonstrate consistent agreement, exhibiting analogous sizes and distributions across all mentioned clusters. This is also the case when examining the compositional loadings (Figure A5), each of which is marked by a clear dominance of type B (60–80%) and type A (10–20%).

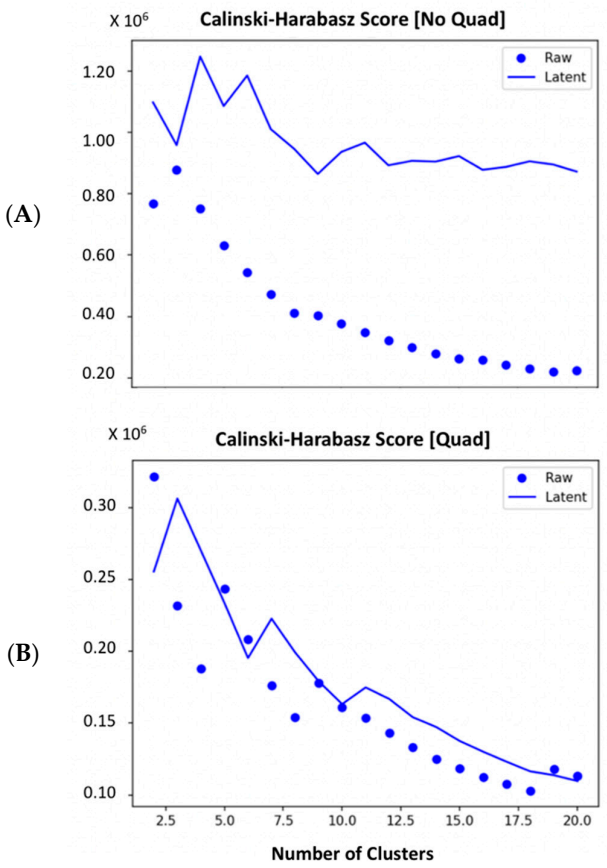

Figure 6. Calinski–Harabasz scores (CH) for the Genie clustered dataset performed on both the unaltered single particle data ('Raw') and the encoded data ('Latent'). (**A**): CH-Score for the data clustered without quadrant information (no quad). (**B**): CH-Score for the data utilising quadrant information (quad).

Contrary to the clear differences presented above, comparisons between the latent and raw genie output are more ambiguous. Qualitatively, the latent heatmap (Figure A4A) displays characteristics consistent with autoencoded data compared to Figure A4C. In particular, Figure A4A displays greater clarity and smoothness as well as maintaining

greater edge consistency compared to the raw output, pointing to probable noise reduction. This is especially clear when comparing identified events, which are clearly defined by the latent output and somewhat amalgamated for the raw data. However, despite the CH-Score describing a large drop-off in cluster quality for the raw data (which is somewhat consistent with qualitative comparisons), Figure A4C does identify additional events on 9 and 10 June. The differences exhibited in the CH-Score (Figure 6A) may in fact highlight some of the challenges presented with applying metrics to encoded data. Specifically, the CH-Score is calculated based on the distances between data points and is applied to the transformed latent space as opposed to the original data. Furthermore, the score assumes an underlying cluster structure, which if not well suited for clustering or exhibiting complex patterns, may not accurately reflect the quality of the clustering solution. Thus, comparing outputs from a reduced dimensionality latent space to the more complex original data may create a natural bias in the score, highlighting the ongoing need for manual interpretation of clusters in addition to applied metrics such as the CH-Score.

3.3. Ambient Airborne Concentration Analysis

Throughout the study period, the 1 min average number concentration of fluorescent particles was 0.38 ± 0.28 cm^{-3}. Of the seven fluorescent particle types, the most abundant were A (average concentration of 0.06 cm^{-3}), B (0.16 cm−3), BC (0.05 cm^{-3}), and ABC (0.05 cm^{-3}). As shown in Figure 7B, the measured particle type was largely size-dependent with A and B corresponding to the smallest particles on average (0.87 and 0.86 µm in diameter, respectively), followed by BC (1.28 µm) and ABC (2.18 µm). The greatest concentration was observed during the first week of the study period on 12 June at 00:33 when the average concentration of fluorescent particles reached 5.1 particles cm−3 (Figure 7C). During this time, the majority (84%) of fluorescent particles were submicron, averaging 0.78 µm in diameter, with a significant portion (54%) attributed to type B fluorescence. Similar events were observed on 6 June at 20:07 and on 8 June at 21:28 when the fluorescent concentrations reached 4.0 and 3.4 particles cm^{-3}, respectively. Of the fluorescent particles measured during the 6 June event, 92% were submicron and 59% were type B. A similar relationship was observed for the 8 June event where 94% of fluorescent particles were submicron and 72% were type B. These three events represent a common occurrence that was observed throughout the study period. Specifically, fluorescent particles peak in the evening as the relative humidity increases. Notably, the higher fluorescent concentration observed on 12 June occurred with a higher relative humidity (83%) when compared to the 6 June (59%) and 8 June (62%). Together, these three events indicate a positive relationship between relative humidity and nighttime fluorescent concentrations and highlight the significant influence of submicron and type B fluorescent particles.

The identification of distinct clusters during this study was found to align with the observed events depicted in Figure 7A. Further examination of the observed events using heatmaps, both for latent and raw Genie analysis (Figure 7A), provided additional insights into the disparities between the two approaches. Interestingly, the raw approach exhibited a lack of clear identification of events observed in June, with a primary focus on events occurring in July. However, this observation aligns with previous comparisons as the two clusters exclusively identified by the raw Genie analysis (refer to Section 3.2) exhibited peak activity between the 5 and 9 of July, as well as the 13 and 14 of July. Although the analysis of these trends shall be reserved for a future study, a note of particular significance is the fact that these periods of peak activity coincided with observed rainfall and a notable increase in the compositional loading of type ABC (Figure A6). Under these specific conditions, coupled with the prevalence of larger-sized particles measuring 3 and 4 µm in relatively low quantities, one can speculate that this activity may be attributed to either fragmented pollen or fungal spores, as observed in a recent study [56]. In contrast, the four clusters exclusively identified through the latent GenieClust analysis were predominantly observed in June, aligning with the peak on 12 June, whilst also exhibiting some activity at the beginning of July.

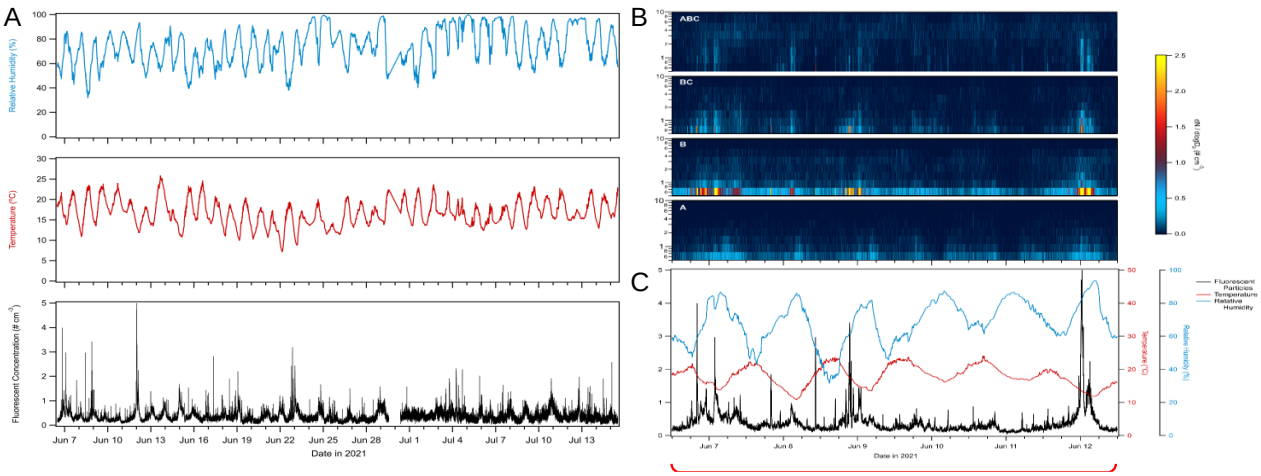

Figure 7. (**A**–**C**): (**A**) (**Left**)—Time series illustrating the fluorescent concentration throughout the duration of the study as well as the corresponding temperature and relative humidity at the supersite. (**B**) (**Top Right**)—Size distribution heatmap displaying particles by their category. (**C**) (**Bottom Right**)—FL conc. Time series (including meteorological conditions) for the first week of the study zoomed in.

4. Discussion

As a study, the breadth of which was to compare the clustering capabilities of various new and traditional approaches, several key findings were made.

The consistency between the transformed clusters identified by K-means and HCA, except for the outlier cluster, highlights the reliability of K-means in capturing the essential patterns within the data. It is important to note that despite the slight discrepancy in the number of clusters identified by K-means and HCA, the overall agreement in cluster profiles and sizes reinforces the effectiveness of both methods in clustering and identifying distinct particle types.

The utilization of an AE in conjunction with the two methods yielded intriguing findings, providing insights into the potential effects on outlier detection for K-means and HCA. Curiously, the incorporation of an AE in the K-means algorithm unveiled differences in outlier detection. By leveraging the capabilities of the AE to capture and represent underlying patterns, K-means may have benefited from the enhanced feature representation by identifying outliers within the main body of the dataset (below 2 microns in this case), particularly in identifying the compositional loadings of each cluster. This enhanced performance appears to be in agreement with a study comparing K-means to AE K-means, which reported an enhanced ability of the autoencoded approach to deal with high-dimensionality data [57]. This was similarly the case for the autoencoded HCA analysis; however, the application of the AE revealed a potential reduction in the number of clusters identified. This appears to suggest the possibility of distortions in important features during the encoding process, leading to a compromised representation of the data and subsequently affecting the clustering outcomes. Furthermore, the observed reduction in clusters implies that the AE may not have adequately preserved the critical information required for HCA to effectively identify and differentiate clusters.

Remarking upon the comparative analysis of GenieClust with K-means and HCA to the transformed dataset, it is clear that GenieClust demonstrated a distinguished ability to detect outliers in this study, challenging the capabilities of the traditional methods. For example, the algorithm successfully distinguished four outlier clusters characterized by larger sizes and lower concentrations, compared to three for HCA and two for K-means. As is evident from the removal of larger particles by the autoencoded analysis, the majority of particles (approx. 80%) captured throughout the campaign appear to lie within the 0.5 micron to 2 micron range. This is of consequence since the combination of the autoencoder with GenieClust yielded unexpected improvements. The AE GenieClust

approach provided the highest number of optimal clusters; however, the most overlapping clusters were identified here also.

Nevertheless, similar to K-means, the AE analysis identified a greater number of distinct clusters in the smaller size ranges. This is likely the consequence of the autoencoder filtering out the larger and less frequent particle data, resulting in the identification of additional distinct clusters within the lower size ranges. However, as a result, the algorithm was unable to identify the three larger-sized clusters previously identified in the transformed analysis.

The findings here appear to align with the existing literature assessing GenieClust's performance. Benchmarks posted by Gagolewski (the algorithm's developer) to the online documentation for GenieClust compare Genie's performance to several agglomerative linkage methods. The Adjusted Rand (AR) Index value for GenieClust was reported to be 3.5, among the lowest mean ranks, suggesting that, on average, GenieClust performs better than many of the other listed algorithms [58].

The study also utilized the Calinski–Harabasz score to compare GenieClust approaches and explored the potential improvement in clustering by including quadrant detector measurements. The optimal number of clusters demonstrated by the score was in good agreement with the manual interpretation, with the raw analysis highlighting 8 clusters and the AE analysis showing 11 clusters. Furthermore, the CH-Score analysis indicated that encoding the data into a latent vector resulted in an increase in cluster quality, as was demonstrated by the consistently higher scores. However, as has been shown here, care must be taken when drawing conclusions from comparative analysis between latent vectors and raw data, owing to the changes in the data structure and how metrics such as the CH-Score determine cluster quality. Additionally, despite a notable decrease in cluster quality exhibited with the inclusion of quadrant information for the CH-Score, the optimal clusters for both demonstrated good agreement, suggesting that the inclusion or exclusion of QI led to relatively small differences and did not provide better cluster separation.

What is evidenced by this study is that GenieClust exhibits a distinguished ability to capture finer patterns and differentiate complex compositions within clusters. Whilst the benefits of this have already been remarked upon, misinterpreting the enhanced precision of GenieClust could lead to an underestimation of the intricacies and multifaceted nature of bioaerosol emissions. In turn, traditional methods, including K-means and HCA, might fail to capture the same level of complexity, potentially leading to the oversimplification of particle associations. Moreover, the integration of AEs alongside other analytical methods introduces the potential for altering the attributes of identified clusters due to dimensionality reduction, which is evident in the divergence between transformed and autoencoded cluster profiles. This aspect holds significant importance, as a failure to acknowledge that the utilization of autoencoders can induce modifications in cluster characteristics could lead to erroneous inferences regarding bioaerosol emission patterns. It is essential to recognise that the application of AEs might bring about shifts in cluster attributes. Overlooking this phenomenon could lead to misconceptions about the advantages of the differences observed following AE implementation. Such an oversight might lead to the dismissal of crucial insights stemming from the filtering effect observed in larger clusters, as demonstrated in this study. As a result, if the methodology were to exclusively rely on AEs, it might entail the inadvertent loss of pivotal information pertaining to bioaerosol emissions.

Limitations and Further Work

Whilst applying an AE yielded mixed results, there are many instances of AEs producing favourable results. Rather, it highlights the complexity of applying such ML techniques compared to traditional clustering methods. Specifically, it is clear that whilst these emerging methods could offer a faster time-to-solution, reduced processing requirements, and the possibility of improved clustering, the complexity of the techniques is significantly increased and requires a fundamental understanding of neural networks. Furthermore, the results here highlight the intricate relationship between the AE, clustering algorithms, and the preservation of crucial data characteristics.

Moving forward there are a number of methods that can address the similarities observed with the autoencoded (latent) clusters:

1. Review and optimize the architecture and hyperparameters of the AE.
2. Applying dimensionality reduction techniques.
3. Explore different AE variations or regularization techniques to enhance the quality of the encoding.
4. Evaluate the clustering performance using various evaluation metrics beyond similarity, such as the silhouette score or cluster purity.

To gain a deeper understanding of the effects observed between the autoencoded K-means and HCA algorithms, comprehensive evaluation measures such as validity indices could be employed to quantitively assess the impacts of the AE on the clustering performance. Furthermore, an examination of the reconstructed data from the AE can provide valuable insights into the distortions and potential loss of information.

Another limitation that should be acknowledged is the absence of training or laboratory data in the current study. This limitation applies not only to GenieClust but also to the K-means and hierarchal clustering analysis conducted. Although it has been observed that GenieClust demonstrates a superior ability to distinguish a higher number of distinct clusters, the lack of reliable training data poses a significant challenge. In order to overcome this limitation and enhance the accuracy of cluster identification, efforts must be directed towards building a comprehensive and reliable repository of known particles. Furthermore, it is important to note that the current limitations in training data extend beyond the scope of this specific study. As researchers aim to delve deeper into the speciation and identification of individual subspecies within broad categories, the lack of reliable training data becomes increasingly apparent. Therefore, there is a pressing need to prioritize the collection and curation of high-quality training data to address this limitation and advance our understanding of complex particle systems.

5. Conclusions

To conclude, this study aimed to assess and compare the clustering capabilities of both traditional and emerging methods, shedding light on their respective strengths and limitations. The consistency observed between transformed clusters using K-means and HCA underscores their reliability in identifying essential data patterns, and whilst HCA did outperform K-means, it is debatable whether the difference seen here were significant enough to warrant the additional processing power and running costs required to run HCA on large datasets. The application of an autoencoder in this study, much in the same way as with GenieClust, appears to have removed infrequent particle data (in this case characterized by larger sizes and low concentrations), the implication being that this may allow for further outlier discrimination within a dataset if used in conjunction with the standard approach. However, GenieClust emerged as a standout alternative, showcasing superior outlier detection compared to conventional techniques. GenieClust's ability to discern finer patterns and its capacity to identify a greater number of optimal and distinct clusters exemplify its enhanced performance. This study's findings have important implications for various domains, including health and pollution monitoring where GenieClust's ability to discern complex particle distributions offers more insightful interpretations. By leveraging GenieClust's strengths, researchers and practitioners can potentially enhance the accuracy and flexibility of unsupervised classification methods. Additionally, the study's exploration of the AE's impacts further contributes to the evolving landscape of data-driven analysis. It is worth noting that future applications could encompass monitoring air quality, tracking pollution sources, and assessing health risks associated with specific particle compositions. The ability to distinguish intricate particle associations using GenieClust holds promise for more effective identification of bioaerosols and pollutant sources and their potential health implications. As technology advances and datasets expand, GenieClust's capabilities may prove an effective alternative to methods that are now beginning to become outdated.

Author Contributions: The following author contributions are acknowledged below: Conceptualization, M.A.N.M., D.O.T., M.W.G., and I.C. built the conceptual sampling approach used in this study; methodology, M.A.N.M. and D.O.T. designed the research methodology and workflow presented here; Data Procurement, M.A.N.M., I.C., and M.J.F. managed data collection and instrument maintanence; software, D.O.T. wrote and applied the machine learning and analysis scripts applied during this study; formal analysis, M.A.N.M. aggregated the outcomes of the tools applied; writing—original draft preparation, M.A.N.M.; and writing—review and editing: M.A.N.M., M.W.G., and D.D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was part funded by the Engineering and Physical Sciences Research Council grant number (EPSRC, EP/S023593/1). This research was also part funded by Droplet Measurement Technologies LLC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be provided via the CEDA web site: https://data.ceda.ac.uk/badc/osca (accessed 4 September 2023).

Acknowledgments: M.M. acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC, EP/S023593/1). M.M. acknowledges funding from Droplet Measurement Technologies LLC.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

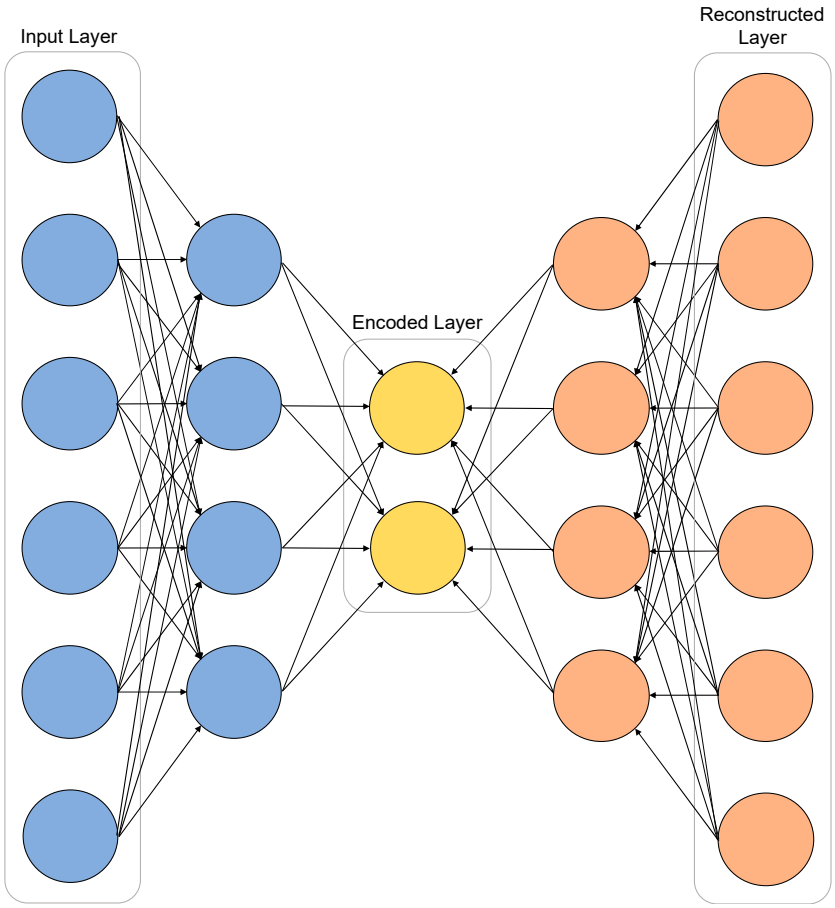

Figure A1. Basic illustration of a neural network/autoencoder architecture.

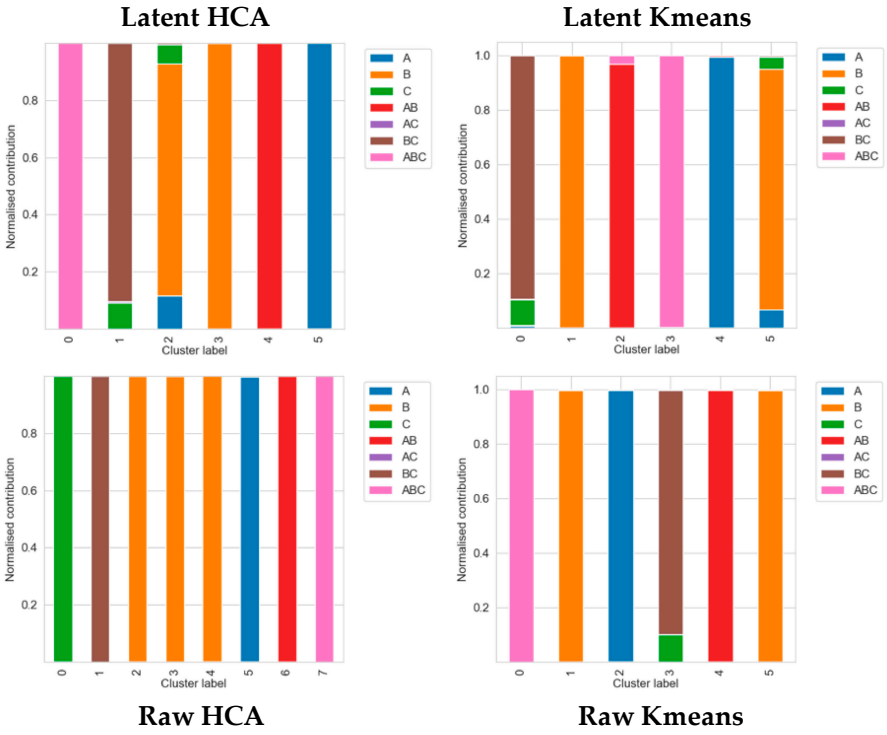

Figure A2. Compositional loading contributions of each cluster identified in the four methods.

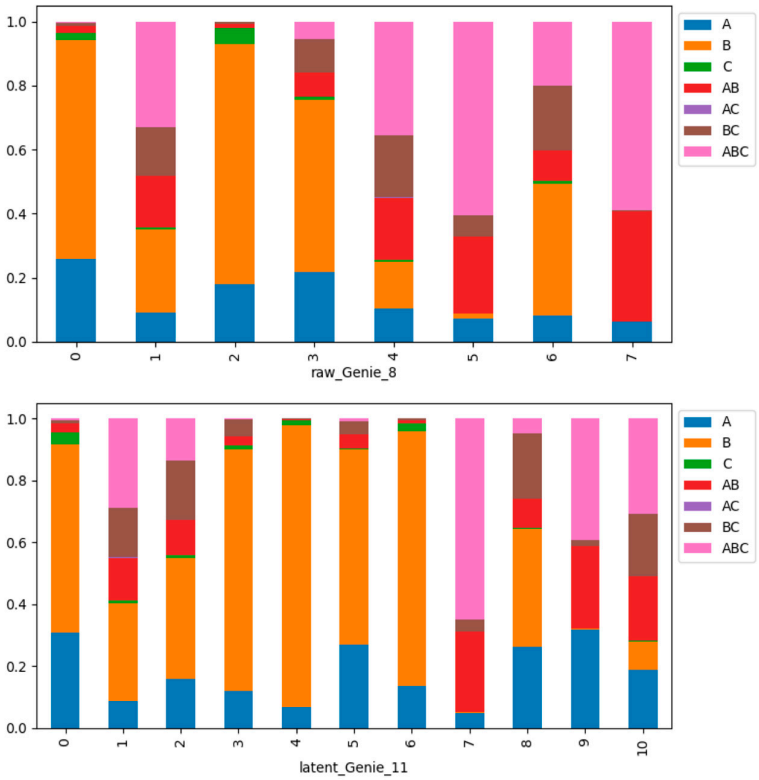

Figure A3. Compositional loading contributions for the two GenieClust approaches corresponding to the number of optimal clusters identified.

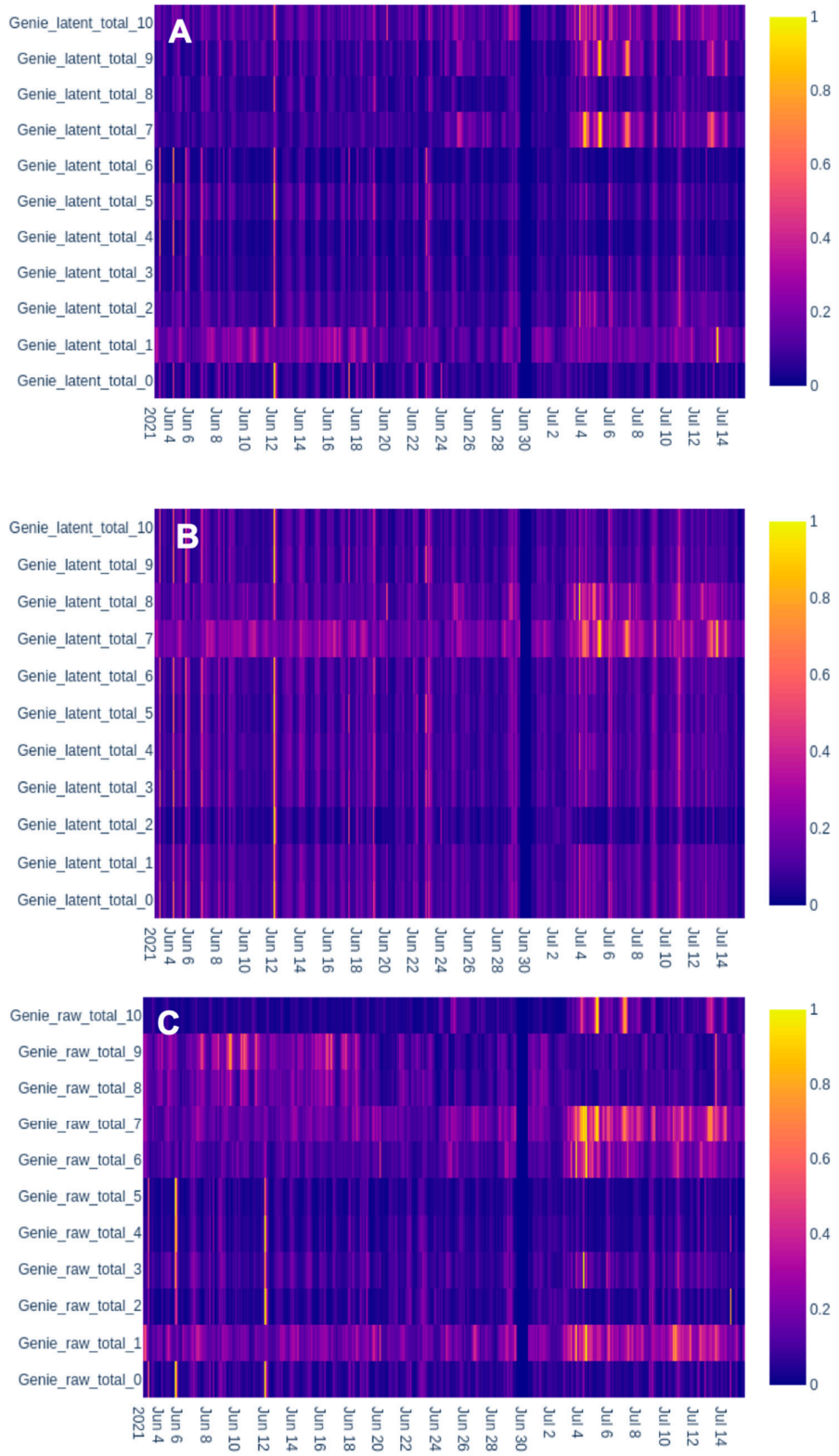

Figure A4. (**A**–**C**): GenieClust heatmaps highlighting cluster events over the course of the study for the AE approach (with and without the quadrant information included), in addition to the transformed approach. (**A**) Latent GenieClust (Without Quadrant Information). (**B**) Latent GenieClust (With Quadrant Information). (**C**) Transformed GenieClust approach (without Quadrant Information).

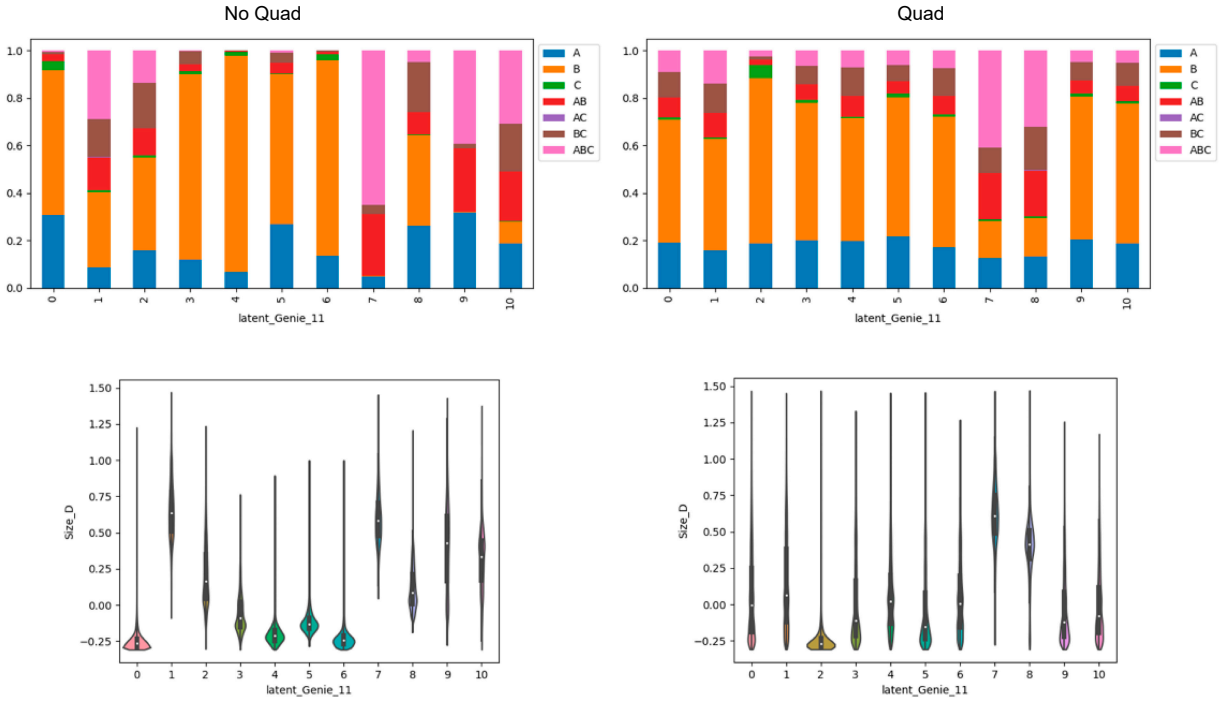

Figure A5. ABC analysis displaying the compositional loadings of each cluster, and violin plots illustrating normalised size distributions for the GenieClust output without the inclusion of quadrant information (**left column**), and with the quadrant information (**right column**).

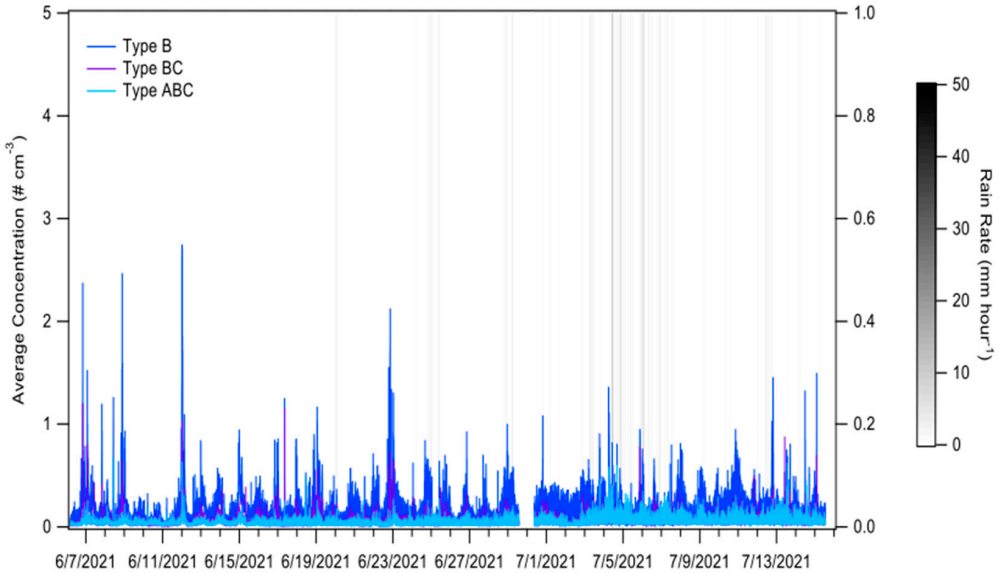

Figure A6. Time series displaying the rainfall and average fluorescent concentrations over the duration of the study for the three most abundant particle categories.

References

1. Patel, T.Y.; Buttner, M.; Rivas, D.; Cross, C.; Bazylinski, D.A.; Seggev, J. Variation in Airborne Fungal Spore Concentrations among Five Monitoring Locations in a Desert Urban Environment. *Environ. Monit. Assess.* **2018**, *190*, 634. [CrossRef] [PubMed]
2. Katz, A.; Alimova, A.; Xu, M.; Rudolph, E.; Shah, M.K.; Savage, H.E.; Rosen, R.B.; McCormick, S.A.; Alfano, R.R. Bacteria Size Determination by Elastic Light Scattering. *IEEE J. Sel. Top. Quantum Electron.* **2003**, *9*, 277–287. [CrossRef]
3. Bradley, R.S. *Paleoclimatology*; Elsevier: Amsterdam, The Netherlands, 2015. [CrossRef]
4. Grgacic, E.V.L.; Anderson, D.A. Virus-like Particles: Passport to Immune Recognition. *Methods* **2006**, *40*, 60–65. [CrossRef] [PubMed]

5. Pearson, C.; Littlewood, E.; Douglas, P.; Robertson, S.; Gant, T.W.; Hansell, A.L. Exposures and Health Outcomes in Relation to Bioaerosol Emissions from Composting Facilities: A Systematic Review of Occupational and Community Studies. *J. Toxicol. Environ. Health Part B Crit. Rev.* **2015**, *18*, 43–69. [CrossRef] [PubMed]
6. Kalogerakis, N.; Paschali, D.; Lekaditis, V.; Pantidou, A.; Eleftheriadis, K.; Lazaridis, M. Indoor Air Quality—Bioaerosol Measurements in Domestic and Office Premises. *J. Aerosol Sci.* **2005**, *36*, 751–761. [CrossRef]
7. Douwes, J.; Thorne, P.; Pearce, N.; Heederik, D. Bioaerosol Health Effects and Exposure Assessment: Progress and Prospects. *Ann. Occup. Hyg.* **2003**, *47*, 187–200. [CrossRef]
8. Huffman, J.A.; Perring, A.E.; Savage, N.J.; Clot, B.; Crouzy, B.; Tummon, F.; Shoshanim, O.; Damit, B.; Schneider, J.; Sivaprakasam, V.; et al. Real-Time Sensing of Bioaerosols: Review and Current Perspectives. *Aerosol Sci. Technol.* **2019**, *5*, 465–495. [CrossRef]
9. Fröhlich-Nowoisky, J.; Kampf, C.J.; Weber, B.; Huffman, J.A.; Pöhlker, C.; Andreae, M.O.; Lang-Yona, N.; Burrows, S.M.; Gunthe, S.S.; Elbert, W.; et al. Bioaerosols in the Earth System: Climate, Health, and Ecosystem Interactions. *Atmos. Res.* **2016**, *182*, 346–376. [CrossRef]
10. Pöhlker, C.; Huffman, J.A.; Pöschl, U. Autofluorescence of Atmospheric Bioaerosols—Fluorescent Biomolecules and Potential Interferences. *Atmos. Meas. Tech.* **2012**, *5*, 37–71. [CrossRef]
11. Wilson, K.H.; Wilson, W.J.; Radosevich, J.L.; DeSantis, T.Z.; Viswanathan, V.S.; Kuczmarski, T.A.; Andersen, G.L. High-Density Microarray of Small-Subunit Ribosomal DNA Probes. *Appl. Environ. Microbiol.* **2002**, *68*, 2535–2541. [CrossRef]
12. Wittmaack, K.; Wehnes, H.; Heinzmann, U.; Agerer, R. An Overview on Bioaerosols Viewed by Scanning Electron Microscopy. *Sci. Total Environ.* **2005**, *346*, 244–255. [CrossRef] [PubMed]
13. Toprak, E.; Schnaiter, M. Fluorescent Biological Aerosol Particles Measured with the Waveband Integrated Bioaerosol Sensor WIBS-4: Laboratory Tests Combined with a One Year Field Study. *Atmos. Chem. Phys.* **2013**, *13*, 225–243. [CrossRef]
14. Song, H.; Marsden, N.; Lloyd, J.R.; Robinson, C.H.; Boothman, C.; Crawford, I.; Gallagher, M.; Coe, H.; Allen, G.; Flynn, M. Airborne Prokaryotic, Fungal and Eukaryotic Communities of an Urban Environment in the UK. *Atmosphere* **2022**, *13*, 1212. [CrossRef]
15. Fennelly, M.; Sewell, G.; Prentice, M.; O'Connor, D.; Sodeau, J. Review: The Use of Real-Time Fluorescence Instrumentation to Monitor Ambient Primary Biological Aerosol Particles (PBAP). *Atmosphere* **2017**, *9*, 1. [CrossRef]
16. O'Connor, D.J.; Healy, D.A.; Hellebust, S.; Buters, J.T.M.; Sodeau, J.R. Using the WIBS-4 (Waveband Integrated Bioaerosol Sensor) Technique for the On-Line Detection of Pollen Grains. *Aerosol Sci. Technol.* **2014**, *48*, 341–349. [CrossRef]
17. Wei, K.; Zou, Z.; Zheng, Y.; Li, J.; Shen, F.; Wu, C.; Wu, Y.; Hu, M.; Yao, M. Ambient Bioaerosol Particle Dynamics Observed during Haze and Sunny Days in Beijing. *Sci. Total Environ.* **2016**, *550*, 751–759. [CrossRef] [PubMed]
18. Gabey, A.M.; Gallagher, M.W.; Whitehead, J.; Dorsey, J.R.; Kaye, P.H.; Stanley, W.R. Measurements and Comparison of Primary Biological Aerosol above and below a Tropical Forest Canopy Using a Dual Channel Fluorescence Spectrometer. *Atmos. Chem. Phys.* **2010**, *10*, 4453–4466. [CrossRef]
19. Petersson Sjögren, M.; Alsved, M.; Šantl-Temkiv, T.; Bjerring Kristensen, T.; Löndahl, J. Measurement Report: Atmospheric Fluorescent Bioaerosol Concentrations Measured during 18 Months in a Coniferous Forest in the South of Sweden. *Atmos. Chem. Phys.* **2023**, *23*, 4977–4992. [CrossRef]
20. Shukla, S.; Naganna, S. A Review on K-Means Data Clustering Approach. *Int. J. Inf. Comput. Technol.* **2014**, *4*, 1847–1860.
21. Singh, K.; Malik, D.; Sharma, N. Evolving Limitations in K-Means Algorithm in Data Mining and Their Removal. *Int. J. Comput. Eng. Manag.* **2011**, *12*, 105–109.
22. Murtagh, F.; Contreras, P. Algorithms for Hierarchical Clustering: An Overview. *WIREs Data Min. Knowl. Discov.* **2012**, *2*, 86–97. [CrossRef]
23. Crawford, I.; Ruske, S.; Topping, D.O.; Gallagher, M.W. Evaluation of Hierarchical Agglomerative Cluster Analysis Methods for Discrimination of Primary Biological Aerosol. *Atmos. Meas. Tech.* **2015**, *8*, 4979–4991. [CrossRef]
24. Tian, J.; Liu, Y.; Zheng, W.; Yin, L. Smog Prediction Based on the Deep Belief—BP Neural Network Model (DBN-BP). *Urban Clim.* **2022**, *41*, 101078. [CrossRef]
25. Yin, L.; Wang, L.; Huang, W.; Liu, S.; Yang, B.; Zheng, W. Spatiotemporal Analysis of Haze in Beijing Based on the Multi-Convolution Model. *Atmosphere* **2021**, *12*, 1408. [CrossRef]
26. Chen, J.; Liu, Z.; Yin, Z.; Liu, X.; Li, X.; Yin, L.; Zheng, W. Predict the Effect of Meteorological Factors on Haze Using BP Neural Network. *Urban Clim.* **2023**, *51*, 101630. [CrossRef]
27. Manimekalai, S.; Prasath, B.; Daniel Shadrach, F.; Lakshmanan, V.; Daniya, T.; Guha, T. Artificial Neural Network with Extreme Learning Machine-Based Wastewater Treatment Systems. In Proceedings of the 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 16–17 October 2022; pp. 1–6. [CrossRef]
28. Shang, K.; Chen, Z.; Liu, Z.; Song, L.; Zheng, W.; Yang, B.; Liu, S.; Yin, L. Haze Prediction Model Using Deep Recurrent Neural Network. *Atmosphere* **2021**, *12*, 1625. [CrossRef]
29. Kwaśny, M.; Bombalska, A.; Kaliszewski, M.; Włodarski, M.; Kopczyński, K. Fluorescence Methods for the Detection of Bioaerosols in Their Civil and Military Applications. *Sensors* **2023**, *23*, 3339. [CrossRef]
30. Xin, Z.; Chen, J.; Peng, H. Advances in Spectral Techniques for Detection of Pathogenic Microorganisms. *Zoonoses* **2022**, *2*, 8. [CrossRef]

31. Markey, E.; Hourihane Clancy, J.; Martínez-Bracero, M.; Neeson, F.; Sarda-Estève, R.; Baisnée, D.; McGillicuddy, E.J.; Sewell, G.; O'Connor, D.J. A Modified Spectroscopic Approach for the Real-Time Detection of Pollen and Fungal Spores at a Semi-Urban Site Using the WIBS-4+, Part I. *Sensors* **2022**, *22*, 8747. [CrossRef]
32. Liu, T.; Duan, F.; Ma, Y.; Ma, T.; Zhang, Q.; Xu, Y.; Li, F.; Huang, T.; Kimoto, T.; Zhang, Q.; et al. Classification and Sources of Extremely Severe Sandstorms Mixed with Haze Pollution in Beijing. *Environ. Pollut.* **2023**, *322*, 121154. [CrossRef]
33. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. *arXiv* **2015**. [CrossRef]
34. Hernandez, M.; Perring, A.E.; McCabe, K.; Kok, G.; Granger, G.; Baumgardner, D. Chamber Catalogues of Optical and Fluorescent Signatures Distinguish Bioaerosol Classes. *Atmos. Meas. Tech.* **2016**, *9*, 3283–3292. [CrossRef]
35. Savage, N.J.; Huffman, J.A. Evaluation of a Hierarchical Agglomerative Clustering Method Applied to WIBS Laboratory Data for Improved Discrimination of Biological Particles by Comparing Data Preparation Techniques. *Atmos. Meas. Tech.* **2018**, *11*, 4929–4942. [CrossRef]
36. Crawford, I.; Gallagher, M.W.; Bower, K.N.; Choularton, T.W.; Flynn, M.J.; Ruske, S.; Listowski, C.; Brough, N.; Lachlan-Cope, T.; Fleming, Z.L.; et al. Real-Time Detection of Airborne Fluorescent Bioparticles in Antarctica. *Atmos. Chem. Phys.* **2017**, *17*, 14291–14307. [CrossRef]
37. Crawford, I.; Lloyd, G.; Herrmann, E.; Hoyle, C.R.; Bower, K.N.; Connolly, P.J.; Flynn, M.J.; Kaye, P.H.; Choularton, T.W.; Gallagher, M.W. Observations of Fluorescent Aerosol–Cloud Interactions in the Free Troposphere at the High-Altitude Research Station Jungfraujoch. *Atmos. Chem. Phys.* **2016**, *16*, 2273–2284. [CrossRef]
38. Watson, N. Meteorological Data from Palas FIDAS 200 Instrument at Manchester Air Quality Site, 2019 Onwards. Available online: https://catalogue.ceda.ac.uk/uuid/62af3c6051044460aa0a716e2204bffc (accessed on 7 August 2023).
39. Forde, E.; Gallagher, M.; Walker, M.; Foot, V.; Attwood, A.; Granger, G.; Sarda-Estève, R.; Stanley, W.; Kaye, P.; Topping, D. Intercomparison of Multiple UV-LIF Spectrometers Using the Aerosol Challenge Simulator. *Atmosphere* **2019**, *10*, 797. [CrossRef]
40. Savage, N.J.; Krentz, C.E.; Könemann, T.; Han, T.T.; Mainelis, G.; Pöhlker, C.; Huffman, J.A. Systematic Characterization and Fluorescence Threshold Strategies for the Wideband Integrated Bioaerosol Sensor (WIBS) Using Size-Resolved Biological and Interfering Particles. *Atmos. Meas. Tech.* **2017**, *10*, 4279–4302. [CrossRef]
41. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
42. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
43. Lieberherr, G.; Auderset, K.; Calpini, B.; Clot, B.; Crouzy, B.; Gysel-Beer, M.; Konzelmann, T.; Manzano, J.; Mihajlovic, A.; Moallemi, A.; et al. Assessment of Real-Time Bioaerosol Particle Counters Using Reference Chamber Experiments. *Atmos. Meas. Tech.* **2021**, *14*, 7693–7706. [CrossRef]
44. Ruske, S.; Topping, D.O.; Foot, V.E.; Morse, A.P.; Gallagher, M.W. Machine Learning for Improved Data Analysis of Biological Aerosol Using the WIBS. *Atmos. Meas. Tech.* **2018**, *11*, 6203–6230. [CrossRef]
45. Forde, E.; Gallagher, M.; Foot, V.; Sarda-Esteve, R.; Crawford, I.; Kaye, P.; Stanley, W.; Topping, D. Characterisation and Source Identification of Biofluorescent Aerosol Emissions over Winter and Summer Periods in the United Kingdom. *Atmos. Chem. Phys.* **2019**, *19*, 1665–1684. [CrossRef]
46. Robinson, N.H.; Allan, J.D.; Huffman, J.A.; Kaye, P.H.; Foot, V.E.; Gallagher, M. Cluster Analysis of WIBS Single-Particle Bioaerosol Data. *Atmos. Meas. Tech.* **2013**, *6*, 337–347. [CrossRef]
47. Fodor, I.K. *A Survey of Dimension Reduction Techniques*; OSTI: Livermore, CA, USA, 2002. [CrossRef]
48. Song, C.; Liu, F.; Huang, Y.; Wang, L.; Tan, T. Auto-Encoder Based Data Clustering. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124. [CrossRef]
49. Chang, C.-P.; Hsu, W.-C.; Liao, I.-E. Anomaly Detection for Industrial Control Systems Using K-Means and Convolutional Autoencoder. In Proceedings of the 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 19–21 September 2019; pp. 1–6. [CrossRef]
50. Guo, X.; Liu, X.; Zhu, E.; Yin, J. Deep Clustering with Convolutional Autoencoders. In *Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 373–382. [CrossRef]
51. Keras-Tuner 1.3.5. Available online: https://pypi.org/project/keras-tuner/ (accessed on 17 June 2023).
52. Zhang, C.; Xia, S. K-Means Clustering Algorithm with Improved Initial Center. In Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, 23–25 January 2009; pp. 790–792. [CrossRef]
53. sklearn.cluster.KMeans. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (accessed on 17 June 2023).
54. sklearn.cluster.AgglomerativeClustering. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering (accessed on 26 June 2023).
55. Gagolewski, M. Genieclust: Fast and Robust Hierarchical Clustering. *SoftwareX* **2021**, *15*, 100722. [CrossRef]
56. Crawford, I.; Bower, K.; Topping, D.; Di Piazza, S.; Massabò, D.; Vernocchi, V.; Gallagher, M. Towards a UK Airborne Bioaerosol Climatology: Real-Time Monitoring Strategies for High Time Resolution Bioaerosol Classification and Quantification. *Atmosphere* **2023**, *14*, 1214. [CrossRef]

57. Wang, X.; Wang, L. Research on Intrusion Detection Based on Feature Extraction of Autoencoder and the Improved K-Means Algorithm. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; pp. 352–356. [CrossRef]
58. Gagolewski, M. Benchmarks (How Good Is It?). Available online: https://genieclust.gagolewski.com/weave/benchmarks_ar.html (accessed on 24 August 2023).