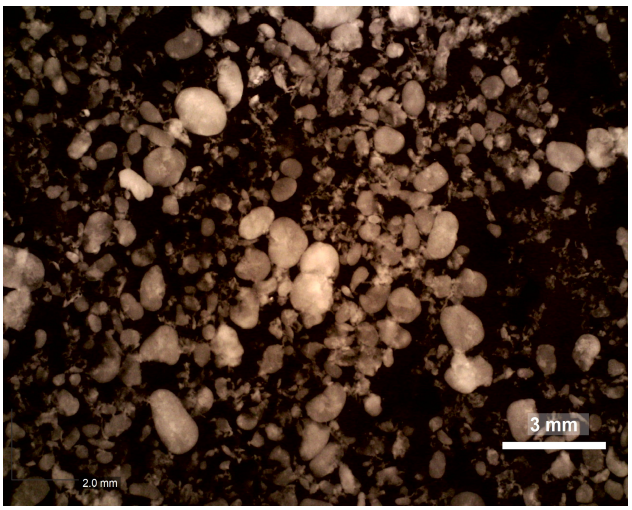
The use of sludge taxonomic profiles to predict the Aerobic Granular Sludge fraction

Report



Author
Mark McCormick

Project
BNF project number 818_2
Analyses multicritères des données des bioréacteurs à boues granulaires pour le traitement des eaux usées
EPFL, Laboratoire de Biotechnologie environnementale, LBE
Station 6, 1015 Lausanne
Chef de projet: Prof. Christof Holliger

Date: September 11, 2023

Contents

1 Abstract

The purpose of this work is to evaluate different methods of using taxonomic count data to predict the sludge fraction of aerobic granular sludge (AGS) samples that were separated by sieving. The two possible sludge fractions considered were 1) granules and 2) mixed flocs/granules (raw, untreated mixed liquor sample). Dimensionality reduction and predictive models were used to identify taxons that are associated with the presence of granules. The PCA, Truncated SVD, logistic regression, Decision tree, Random Forest, and Naive Bayes Classifier methods were assessed. The study data were obtained from laboratory bioreactors during 942 days of continuous operation under different substrate feeding and operating regimes. The prediction dataset was an array of 851 bacteria taxons identified and counted in samples taken on 326 days. The response dataset was a vector of 2 possible response values (0 or 1) that identified the samples as granules separated by sieving or as mixed flocs/granules. The methods developed during this study might be useful for understanding the differences in the taxonomic profiles of AGS granules and mixed floc/granule suspensions.

The conclusions of this study and the corresponding method used are:

- **PCA:** the two different fractions were made apparent using PCA and cluster definition.
- **PCA:** the populations contained in the granules are not so different from the populations that are not in the granules (mixed flocs/granules) (Bray-Curtis index = 0.14).
- **Logistic regression:** the logistic regression function most accurately predicts the samples identified as the granule fraction from raw taxon counts (Prediction accuracy: 90%).
- **Logistic regression:** the presence of *Dechloromonas* is negatively correlated with the granule fraction
- **Logistic regression:** the realtively large positive values of the regression coefficients of *Tahibacter*, *Xanthomonodaceae*, and *Bdellovibrio* suggests that these taxons are positively correlated with the granule fractions.
- **Decision tree:** the presence of *p Verruccomicrobia*, a known polysaccharide degrader, is critical to making the sample classification.
- **Random forest:** the taxons counts of *p Verruccomicrobia*, *g Candidatus Competibacter*, *g Denitratisoma*, and *c WCHB1-32* are critical predictors of granules.
- **Preprocessing:** preprocessing to yield a more normal data distribution does not always improve prediction accuracy. Use of the raw data should be considered.

2 Introduction

Improving solids separation by sedimentation makes it possible to reduce the size and the operating costs of wastewater treatment plants. Aerobic sludge that contains granules has better settling characteristics than aerobic sludge that contains flocs or that does not contain granules. Consequently, the conditions under which Aerobic Granular Sludge (AGS) forms is an important research topic. A summary of the process and of AGS process development can be found in [1].

The microbial community composition of Aerobic Granular Sludge (AGS) depends on multiple intrinsic and extrinsic factors related to the community member functions or to the environment [1]. Consequently, understanding the relation between granule formation and the microbial community composition of individual granules will contribute to successful implementation of the AGS process.

With the goal of contributing to efforts to understand the relation between community composition and granule formation, this report evaluates different numerical methods for predicting AGS mixed liquour sample class (granule or mixed) based on sludge taxonomic profile counts. A separate presentation (Modeling AGS bioreactor experiments) describes the use of machine learning methods to predict AGS formation from the bioreactor operating conditions.

3 Methods

This section briefly describes the methods used. The data sciences tools, the research questions, and the practical use are introduced. They are described in more detail in the results section.

3.1 Data acquisition

All the data used to make this study were from the PhD thesis of Aline Adler [1]. Taxon identity and abundance (counts) were determined in the samples of Aerobic Granular Sludge mixed liquour. A sample consisted of mixed liquor containing between 1 and 2 ml of wet biomass. Samples containing granules were obtained by sieving with a 250 μm cut-off. The samples were centrifuged, washed and homogenized before DNA extraction.

The samples were from long-term laboratory bioreactor cultures operated to investigate the change in microbial community structure as a function of the influent type and the operating conditions. Two separate bioreactors (RA and RB) were operated. The different bioreactor influent types were:

- Simple 1 (C source = VFA)
- Simple 2 (C source = VFA)
- Transition from simple to another type
- Complex monomeric
- Transition from complex monomeric 1 to another type
- Complex polymeric
- Transition from complex monomeric 2 to another type

This study considers only the taxon counts and the sample classification as one of either granules or mixed flocs/granules (raw, untreated mixed liquor). No attempt was made to associate the influent type, the culture conditions, or a sequence of events in time with the presence of granules. Taxon identities and counts were determined using Metagenome-Assembled Genome (MAG) methods. Detailed descriptions of data aquisition can be found in [1] and [2].

3.2 Raw data pre-processing

A new dataframe was created from the experiments conducted between 25.09.2015 to 24.04.2018. The columns were the sample ID (including the sampling date), and added columns corresponding to the sample fraction (granule or mixed flocs/granules). Each dataframe row held the taxon counts and the sample fraction type for a sample. Samples were not collected on all days. On the dates when samples were collected, one sample was collected per day. In summary, the dataset comprised:

- Predictor array: comprised of 851 bacteria taxons identified and counted in samples taken on 326 days
- Response vector: Observed sample fraction: 1) granules separated by sieving or 2) mixed flocs/granules (raw, untreated mixed liquor sample)

As shown in Figure 1, the taxon richness (bar color codes) and abundance (bar heights) varied greatly during the study period.

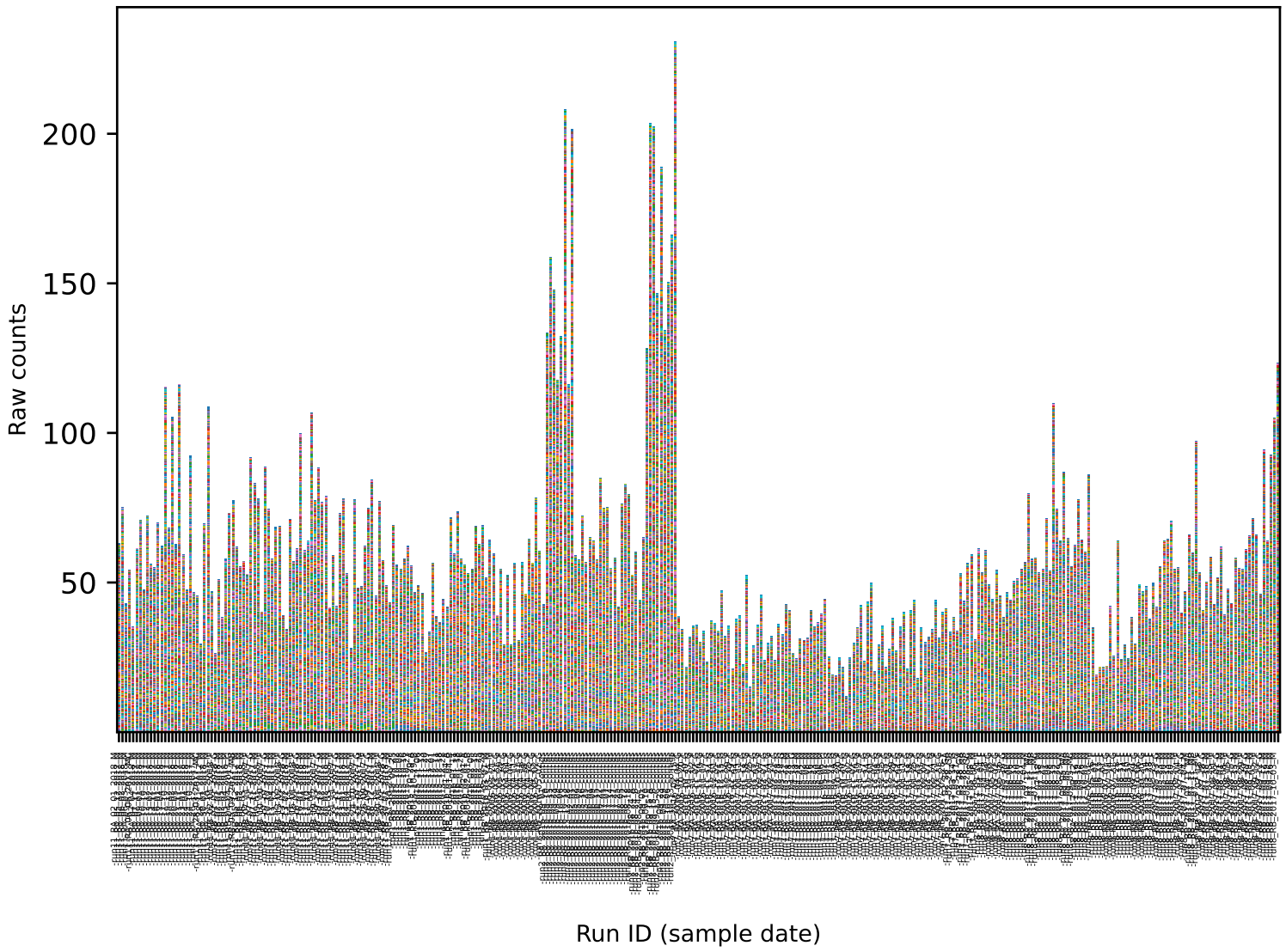


Figure 1: **Raw taxon counts per sample (sample date). Note: the series is not in chronological order**

Non-normally distributed data points are known to reduce the accuracy of the predictions made using machine learning models. The Shapiro-Wilks test [15] was used to evaluate the distribution of the counts of each taxon. The value of the test statistic lies between 0 and 1. A normal distibution has a value of 1. As shown in Figure 2, the value of the test statistic varied between 0.11 and 0.17 for all of the taxons in the raw data. This shows that the taxons counts are not normally distributed. Consequently, preprocessing of the raw data to normalize the distribution should be considered.

A preprocessing method known as Standard scaling is frequently used to transform raw data to obtain a normal distribution. The standard score of a sample is calculated by subtracting the mean value from the samples and dividing by the standard deviation of the samples. As shown in Figure 3, after standard scaling the value of the test statistic varied between 0.4 and 0.9 for all of the taxons. The mean value was 0.67. This shows that data preprocessing using standard scaling resulted in a more normal data distribution. Figure 4 shows that data preprocessing using the MinMax scaler also resulted in a more normal data distribution. The mean value was 0.59. In all cases, the p-value of the test statistic was less than 3.5 x 10^{-16}.

Since high prediction accuracy is the goal of model building, the preprocessing step should be selected based on the accuracy of the predictions made by the model and not on the normality of the data distribution.

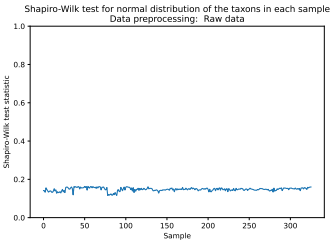


Figure 2: Raw data

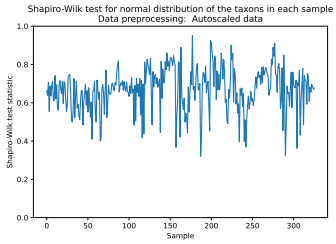


Figure 3: Standard scaled data

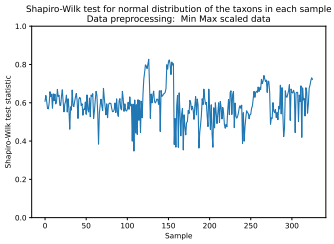


Figure 4: MinMax scaled data

The response vector was created by extracting information from the raw data frame sample IDs. The sample IDs are strings that contain information about the sampling date, run number, reactor ID, and the observation of granules. A new column, labelled 'G_True' contained a 0 or a 1 value depending on, respectively, the presence or the absence of a 'G' in the sample label. The letter 'G' designates granules. The raw, mixed liquor samples did not contain a 'G' in the label. In this way, 2 classes of samples were obtained.

The presence or absence of granules and flocs is shown in the Figure 6. The number of samples based on granule quality classes is:

- samples having granules: 84
- raw, mixed liquor samples: 242
- total samples: 326

The counts are summarized in the Figure 5.

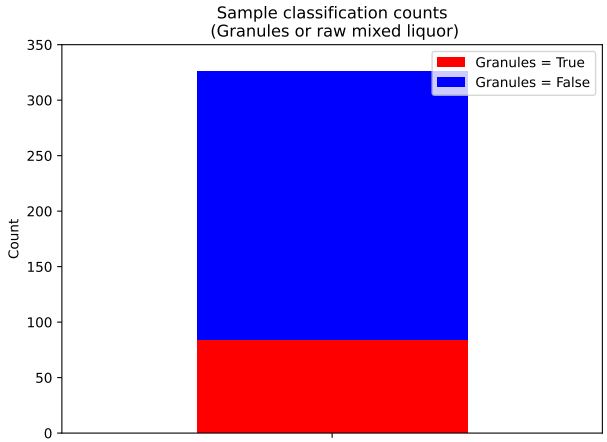


Figure 5: **Proportion of sample types: 1) Granules (Granules = True) and 2) Raw, mixed liquor (Granules = False)**

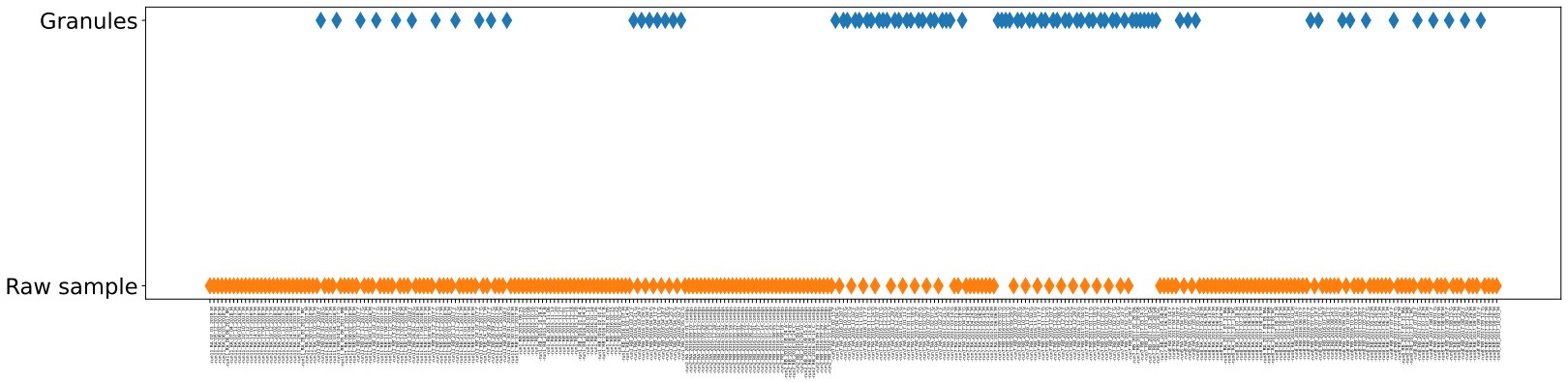


Figure 6: **Response data vector**

3.3 Machine learning

The term "Machine Learning" means that the data are split into 1) a training set that the model uses to learn the relation between prediction and response data, and 2) a test set used to test the accuracy of predictions made from previously unseen data. Many open source libraries of tools are available for use in developing custom pipelines to extract, transform, and load the data into machine learning algorithms and to visualize the raw data and the results. The open source libraries Scikit-learn [7], Scipy [15], and Matplotlib [6] were used respectively for machine learning, data analysis, and visualization.

The quality of a machine learning model is evaluated in terms of the accuracy score. Consequently, the main activities of the data scientist are transformation of the raw data and opimisation of the model parameters in order to improve the accuracy of the predictions. In this study, the machine learning is "supervized" since both the predictors and the corresponding responses are known.

3.4 Principal components analysis (PCA)

Principal components (PC) are obtained from the m x n data matrix where m is the number of rows (observations/sample counts) and n is the number of columns (variables/taxons). One use of PCA is to reduce a large number of measured variables to a smaller number of latent variables called principal components. Each principal component groups correlated data into a single latent variable thereby reducing the dimensionality of the dataset. This numeric method is based on decomposing the data matrix into factors that make it possible to describe the data in multidimensional space with fewer dimensions. The principal axes of this space are obtained from the covariance matrix of the array columns (experiment variables). Information such as the fraction of total variance explained by the principal component, the identity of the highly correlated variables, and the relative importance of each variable in defining the principal component can be extracted from the principal component axes [3].

Before using the PCA algorithm, the raw data is centered at zero by applying a standard scaler function to subtract the mean and divide by the standard deviation. The PCA scores are the distances from the data center point to a perpendicular projection of the raw data point to the PC axis. The scores are the sums of the products of the distance from the center when a sample value is projected to the PC axis and the loading of each variable in a sample. The loadings are the factors that orient the PC in the multidimensional space. The PC axes are calculated to explain as much data as possible. If the variables are highly correlated, then they are reduced to only a few latent variables and the first 2 or 3 PCs will explain most of the data.

The first 2 principal components are often represented as a Bi-plot of 1^{st} component scores versus 2^{nd} component scores. Data points that are clustered together on the plot have similar scores. This often implies that the predictor variables have similar impacts on the sample. The Bi-plot also reveals outliers and distinct clusters.

The data points can be color coded to represent the value of a response variable or another variable of interest [4]. In this study, the data set is an m x n matrix where m is 326 samples taken almost always on different dates and n is 851 different taxa searched for and counted in each sample. The data points are color coded to show the presence or the absence of granules in the sample.

Practical use: Samples of granules observed on the Bi-plot can be grouped into separate clusters. Although the cluster limits are defined by the observation of granules, in most cases some samples without granules will also be observed inside the cluster limits. The taxonomic profiles, as descibed by taxon counts, of the samples inside the cluster are expected to be different from the taxonomic profiles of samples that are outside of the cluster.

The Bray-Curtis index [13] shown in Equation (1) was used to evaluate the difference in taxon counts between the population inside the cluster and the population outside of the cluster. A value of 0 means that the population inside the cluster and the population outside the cluster are the same in terms of taxons and counts. A value of 1 means that the populations are different.

$$BC = \frac{\sum |u_i - v_i|}{\sum |u_i + v_i|} \tag{1}$$

where u_i is the taxon count outside the cluster and v_i is the taxon count inside the cluster.

3.5 Truncated Singular Value Decomposition (SVD)

As shown above, the raw taxon counts are not normally distributed. In general, data preprocessing for machine learning includes methods to normalize and center raw data. However, data normalization reduces the importance of high counts, increases the importance of low counts and consequently might remove useful information. In a way similar to PCA, Singular Value Decomposition (SVD) is a numeric method that decomposes the data matrix into factors that make it possible to describe the data in multidimensional space. Truncated SVD is a method specifically developed for sparse matrices such as word counts [9]. The data is not centered and all but the first principal components are set to zero. Since the taxon count matrix is a sparse matrix of counts, the truncated SVD method might yield more accurate predictions than PCA.

Practical use: The dataset used in this study is a sparse matrix of non-normally distributed counts. Consequently, the type of data transformation to use in an important question. The dimensionality reduction using the truncated SVD method on raw data is expected to achieve more dimensionality reduction than using the PCA method. The dimensionalty reduction can be evaluated in terms of the % of variance explained by the first 2 principal components.

3.6 Logistic regression

Logistic regression is a linear model for classification [10]. It is a special case of a linear model where the probability of the response variable has one of only two possible values. In this study, the use of logistic regression is appropriate because the response variable is not a continuous variable, but rather, a discrete variable (Granules or mixed flocs/granules raw mixed liquor).

Practical use: Use of the Logistic regression method might achieve higher predicion accuracy than use of other methods (Decision tree, Random forest, Naive Bayes classifier). The prediction accuracy can be evaluated in terms of the score. The score is the fraction of samples with a correct predictions of granules.

The relative values of the coefficients of the logistic regression model indicate the relative importance of the corresponding predictors. This information can be used to identify the taxons that are characteristic of granules in an AGS producing culture.

3.7 Decision Tree

A decision tree is a supervized learning method that predicts the value of a response variable by learning decision rules inferred from the predictor values [8]. The model creates a tree of branching nodes that present a question that results in sample classification into 2 bins (branches). The model parameters are set to maximize the accuracy, generalization, and interpretability of the tree.

Practical use: Use of the decision tree method might achieve higher predicion accuracy than use of other methods (Logistic regression, Random forest, Naive Bayes classifier). The prediction accuracy can be evaluated in terms of the score. The score is the fraction of samples with a correct predictions of granules.

3.8 Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the prediction accuracy and control over-fitting [12]. In this study a Random forest is implemented as a machine learning supervized model to predict granules from sample taxon counts.

Since a Random forest is a collection of many decision trees trained with reinitialized weights, the predictions are made by averaging the results. Consequently, they are more accurate than those made using a single decision tree.

A confusion matrix is used to evaluate the prediction accuracy in terms of True positives, True negatives, False positives, and False negatives where positive = granules, and negative = raw, mixed liquor.

Practical use: Use of the Random forest method might achieve higher predicion accuracy than use of other methods (Logistic regression, Decision tree, Naive Bayes classifier). The prediction accuracy can be evaluated in terms of the score. The score is the fraction of samples with a correct predictions of granules.

3.9 Naive Bayes Classifier

Naive Bayes classifiers are supervized learning algorithms that apply Bayes' theorem with the naive assumption of independance between every pair of features. Bayes' theorem uses conditional probability calculated from data to predict the probability of a future event. Bayes' theorem states the following relationship between events and probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

where event A is granules obtained by sieving and event B is the taxon count.

The multinomial Naive Bayes classifier implements the naive Bayes algorithm for classification of multinomially distributed data with discrete features such as feature counts [11]. A confusion matrix is used to evaluate the prediction accuracy in terms of True positives, True negatives, False positives, and False negatives where positive = granules and negative = raw, mixed liquor.

Practical use: Use of the Multinomial Naive Bayes classifier method might achieve higher predicion accuracy than use of other methods (Logistic regression, Decision tree, Random forest). The prediction accuracy can be evaluated in terms of the score. The score is the fraction of samples with a correct predictions of granules. By distinguishing True and False postives and negatives, the confusion matrix gives more detail about the accuracy.

4 Results and Discussion

The results for dimensionality reduction and machine learning techniques are presented below.

4.1 Principal components analysis (PCA)

Principal components analysis was used for dimensionality reduction. As described above, 2 qualities of AGS samples were considered: Granules and raw, mixed liquor. Additionally, the preprocessing of raw data to remove outliers and make a more normal distribution of the data was considered.

Consequently, the following cases were evaluated:

- Raw data
- Standard scaled data
- MinMax scaled data

The hypothosis is that samples of granules can be grouped into separate clusters and the taxonomic profiles, as descibed by taxon counts, of the samples inside the cluster are different from the taxonomic profiles of samples that are outside of the cluster. This hypothesis can be tested using the Bray-Curtis index.

The Bray-Curtis index [13] shown in Equation (1) was used to evaluate the difference in taxon counts between the population inside the cluster and the population outside of the cluster. A value of 0 means that the population inside the cluster and the population outside the cluster are the same in terms of taxons and counts. A value of 1 means that the populations are different.

The clusters were centered on the PC1 and PC2 median values. The cluster limits (green line in the figures) were defined arbitrarily as a coefficient X the standard deviation of PC1 and a coefficient X the standard deviation of PC2. Some of the samples lying inside the cluster limits are not samples of granules.

Test method The sample dataset was visualized as a Bi-plot of PC-1 and PC-2 scores. Each point on the bi-plot represents the score of a single observation (sample date). The points were color coded to show the granule samples and the raw, unfiltered mixed liquor. The count of each of the 851 taxons contributes to the score value. Each taxon has a corresponding PC loading. Consequently, samples that have similar PC scores have similar taxon counts and contributions to the orientation of the PC in multidimensional space.

The method is summarized as follows:

1. Make a Bi-plot of PC-1 vs PC-2 scores.
2. Color code the samples according to granules or raw, mixed liquor.
3. Based on inspection of the Bi-plot, define clusters of samples that have granules.
4. Evaluate the Bray-Curtis index of the taxon counts of the populations inside and outside of the cluster.

In addition to the use of PCA to identify clusters, the use of PCA for dimensionality reduction was evaluated. In the case of analysis of the raw data from all samples, the first 2 principal components explain respectively 68.1 % and 15.3% of the data (total = 83.4%).

The results of the 3 data treatments evaluated (raw data, standard scaling, MinMax scaling) are presented in Figures 7, 8, and 9 below.

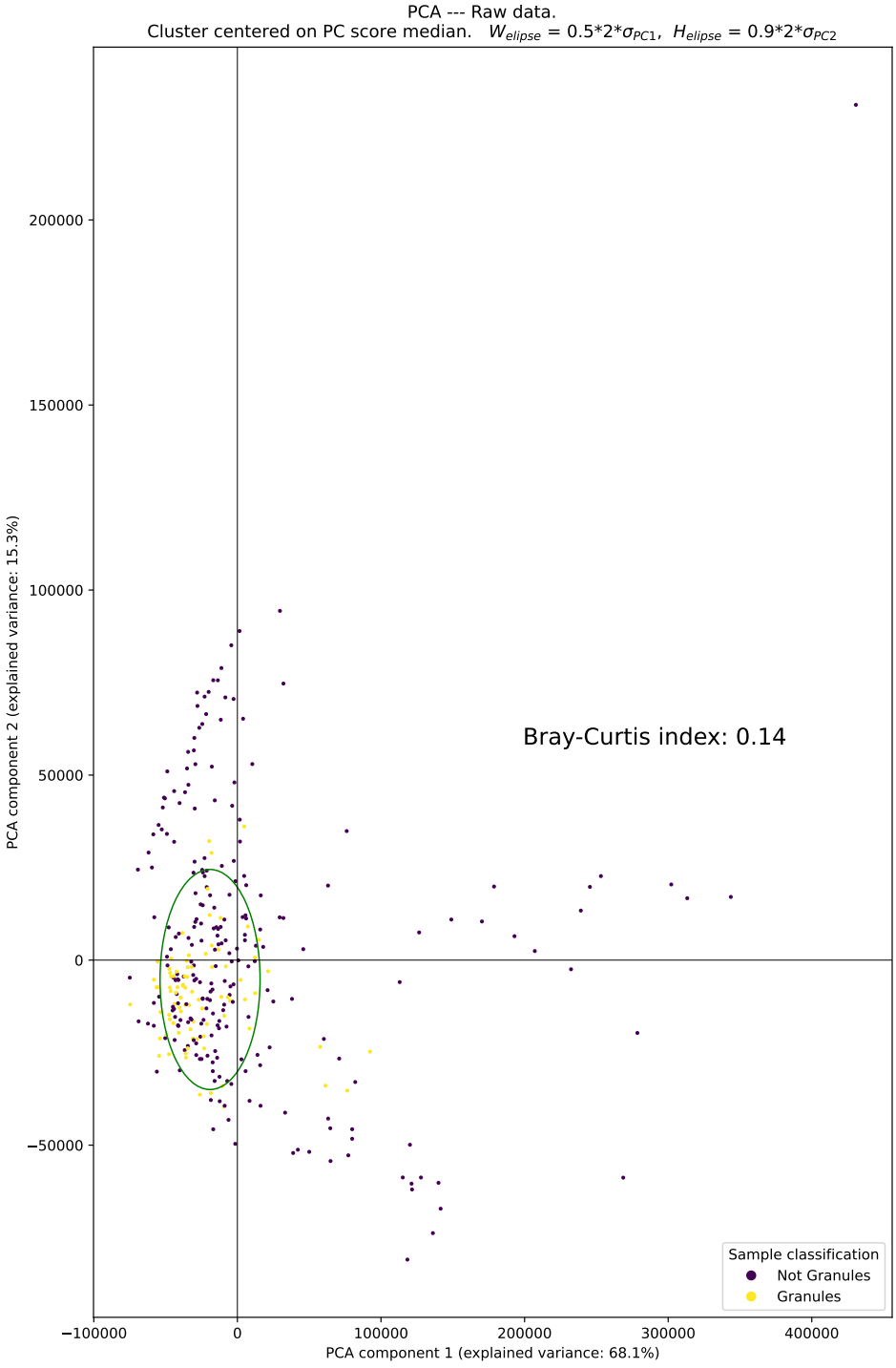


Figure 7: **Taxon bi-plot with granule cluster, raw data**

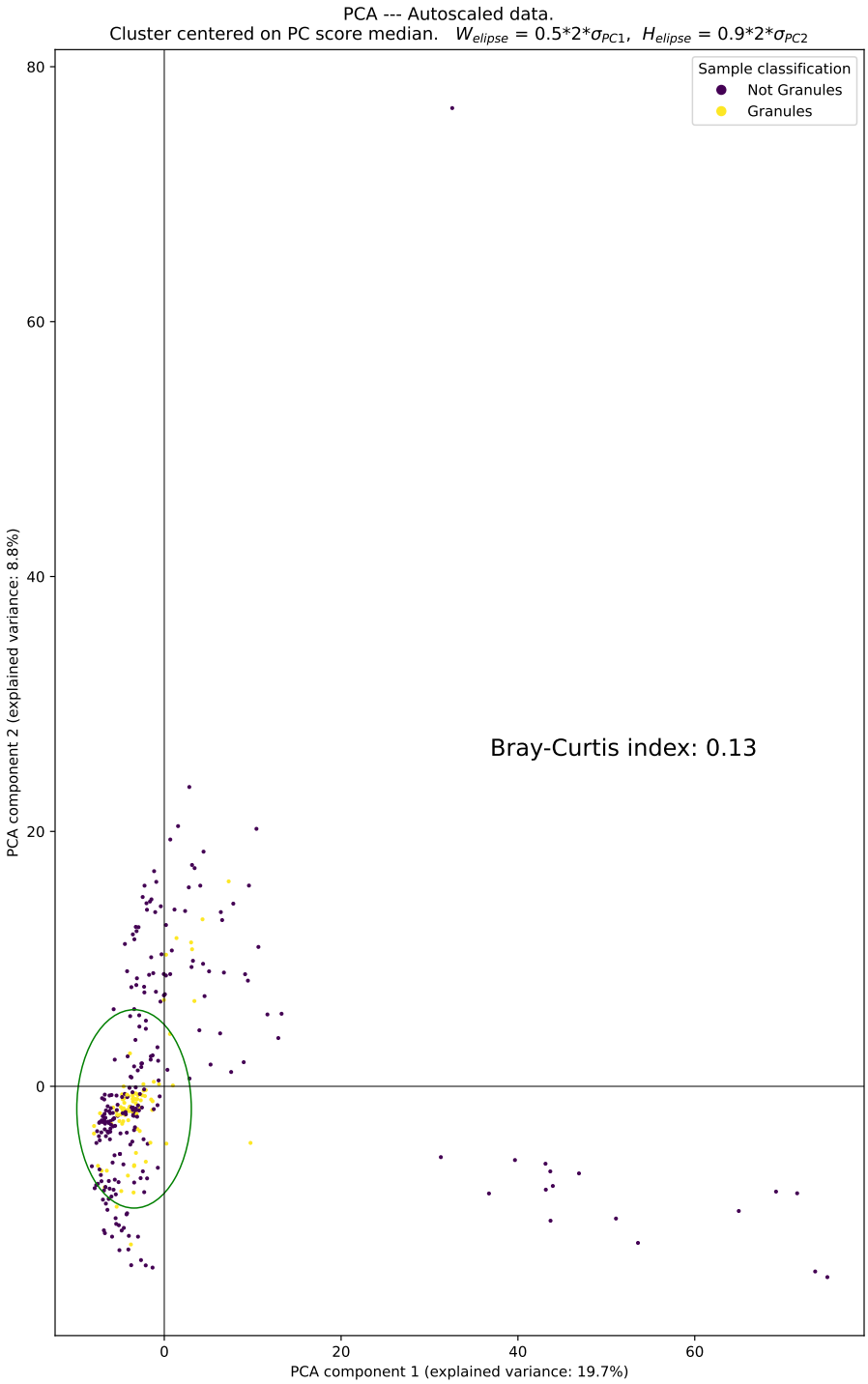


Figure 8: **Taxon bi-plot with granule cluster, standard scaled data**

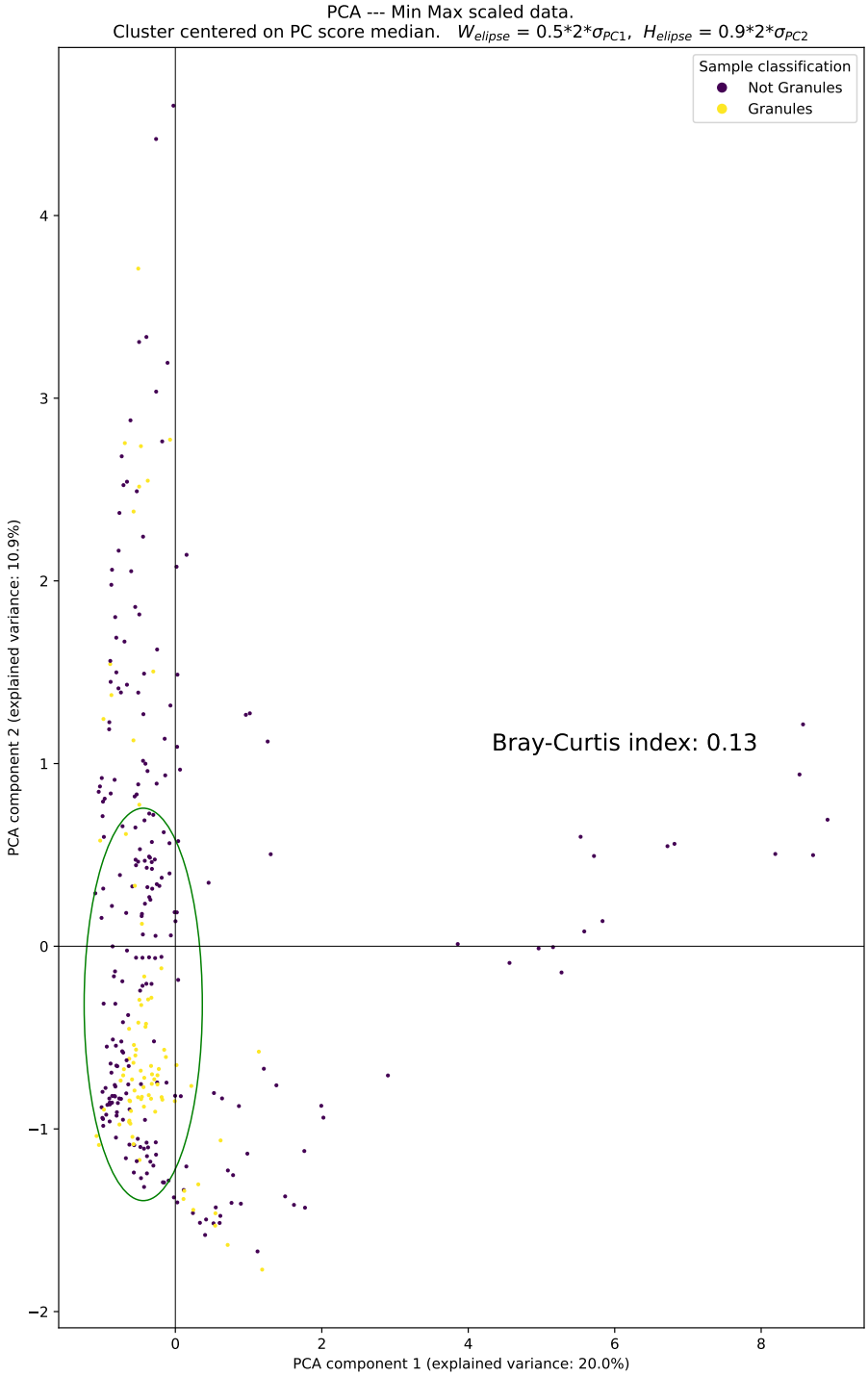


Figure 9: **Taxon bi-plot with granule formation cluster, MinMax scaled data**

The difference between the populations inside and outside of the cluster was also evaluated in terms of counts of the individual taxons.

Conclusions on cluster analysis

Using raw data and using both of the normalisation methods, a single cluster could be identified based on observation of the color coded samples on the Bi-plot. The Bray-Curtis index was used to compare the distance of the populations inside and outside of the cluster. The highest Bray-Curtis index (0.14) was obtained for the raw data. This low value implies that the population in the cluster of granules is not very different from the population outside of the cluster.

Since PCs reveal correlations, the observation of clusters implies that all the samples located inside the cluster have a special relationship with each other. For example, in the samples inside the cluster limits but not identified as granules, some of the taxons might be contributing to future granule formation. Alternatively, some of these taxons might have contributed to recent granule disintegration or to floc formation. Additionally, samples that are distant from the cluster contain taxons that do not favor granule formation. These taxons might have no effect on granule formation. They might interfere with granule formation. Or, the samples might simply lack the taxons required for granule formation.

The Bray-Curtis indexes are presented in Table 4.1.

Data treatment	Mixed granule samples removed
Raw (no treatment)	0.14
Standard scaled	0.13
MinMax scaled	0.13

The present analysis focussed on mixed liquor samples preprocessed for selection of granules. In future studies, it might be possible to modify the method for use in distinguishing raw mixed liquor that is likely to form granules from raw mixed liquor that is not likely to form granules. This approach might be useful for the operation, trouble shooting, or the optimisation of wastewater treatment plants.

Identification of important taxons

To identify the differences between the poulations inside and outside of the cluster limits, both the taxon loading values and the taxon counts were compared.

The PC loadings orient the PC axis in the multidimensional data space. The loading values result from the dimensionality reduction. This implies that the importance of a particular taxon in defining the PC axis is proportional to its loading value and thus to the distribution of the data.

The ranked loading values reveal the taxons that are the most important in orienting PC1 in the multidimensional data space. This implies that the observed counts of these taxons might have an important incidence on the inter correlation of the data. Based on the higher Bray-Curtis index and the explained variance of PC1 and PC2 shown above, the raw data was used without preprocessing. The top ranked taxons, in terms of PC1 loading values in descending order, are shown in Table 1. Comparison of the loadings might be useful when assessing, troubleshooting and optimising AGS processes.

Table 1: **Top taxons ranked by PC1 loading values**

Taxon ID	Taxon
0.1.1.3.1.2	f B142
0.1.1.1.3.1	f Competibacteraceae
0.1.1.3.1.4	f Meganemaceae
0.1.6.1	c Anaerolineae
0.1.1.3.7.3	f Rickettsiales Incertae Sedis
0.1.1.3.1.5	f Hyphomicrobiaceae
0.1.3.2.1	o Cytophagales
0.1.1.3.1.12.1	g JG35-K1-AG5
0.1.4	p Saccharibacteria
0.1.1.3.1.7.1.1	s unclassified
0.1.1..2.2.1.15	s Ottowia
0.1.8.2.2	o ODP123OB30.09

The differences in the mean taxon counts between all samples inside the cluster limits and all samples outside the cluster limits was assessed. Considering taxon counts, the figures below show the highest ranked taxons in the populations inside and outside of the cluster. Raw data was used without normalization because the use of raw data resulted in a very high explained variance in PC1 and a slightly higher Bray-Curtis index. Figure 10 shows that *g Candidatus Accumulibacter* is the 11^{th} highest ranked taxon in the granules and the 15^{th} highest ranked taxon in the raw liquor. The mean count of *g Candidatus Accumulibacter* in the raw liquor was almost the same as in the granules. This suggests that a high concentration of *g Candidatus Accumulibacter* in the mixed liquor might be required for granule formation.

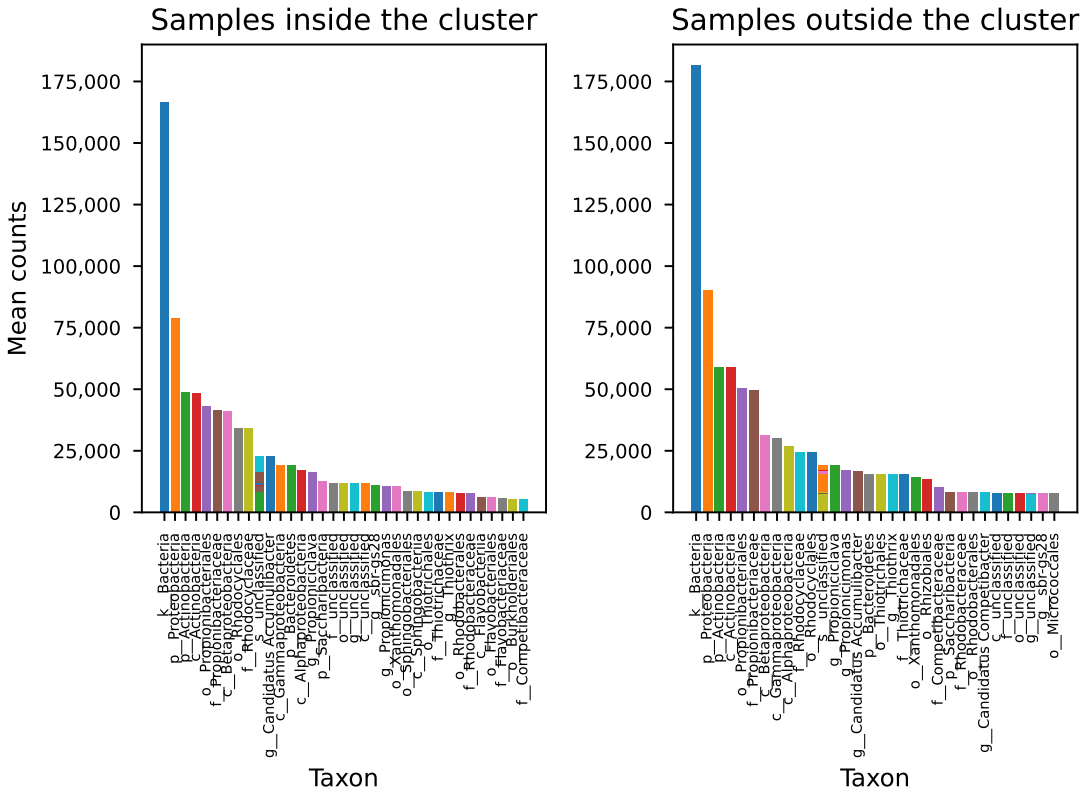


Figure 10: Ranked taxon counts

Figure 11 shows the differences in the taxon counts between samples located inside and outside of the cluster limits. The taxon *g Candidatus Accumulibacter* is more abundant inside the cluster limits than outside the cluster limits. The taxons *c Betaproteobacteria*, *f Rhodocyclales*, *o Rhodocyclales*, and *p Saccharibacteria* are also more abundant inside the cluster limits than outside the cluster limits.

Considering all the classes, *Actinobacteria* has high count values in the samples located outside of the cluster limits. This is consistent with the report that *Actinobacteria* abundance changes during transitions from simple to complex wastewater [2] and the influent feed type transitions of the present study. The taxons *Alphaprotobacteria* and *Rhizobiales* had higher count values in the mixed liquor than inside the cluster limit. As shown in Figure 10, *Gammaproteobacteria* was abundant outside the cluster limits. This is consistent with the results from another study where in an experimental reactor fed synthetic wastewater *Gammaproteobacteria* affiliating *Rhizobiales* from *Alphaprotobacteria* were present in up to 17% of the dense granule community[16]. However, in the present study, these taxons are present with the taxons *Alphaprotobacteria* and *Rhizobiales* more abundant outside than inside the cluster. Also consistent with a previos study, *Propionimonas* has a high count value both inside and outside the cluster with higher counts outside of the cluster. *Propionicimonas* has been found to be enriched in complex wastewaters [2].

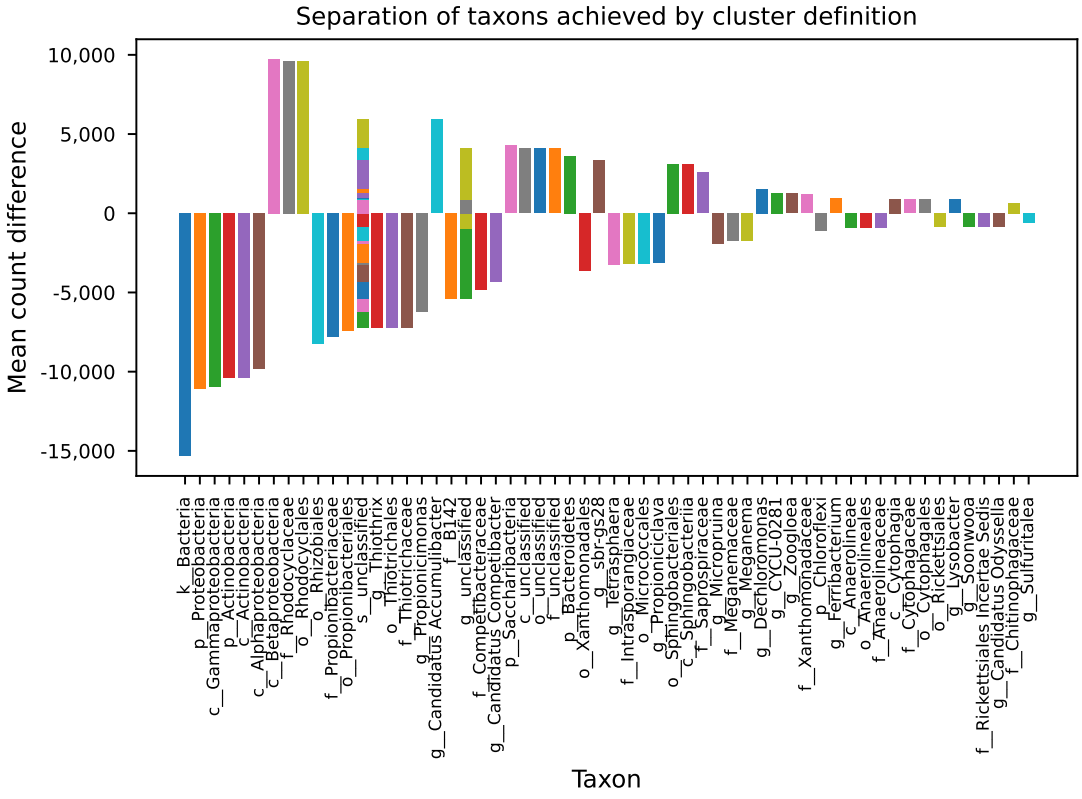


Figure 11: Taxon count differences ranked by absolute value
Mean count difference = Mean count inside cluster limit - Mean count outside cluster limit

Comparison of cluster taxon counts

The taxon counts inside and outside of the clusters created using PCA can also be used to evaluate the granule formation. Figure 12 shows color coded taxon counts classified by cluster and the presence of granules. The figure shows a contrast between yellow and blue in the different classes for some taxons. For other taxons, the contrast is small indicating that the counts are not different between classes.

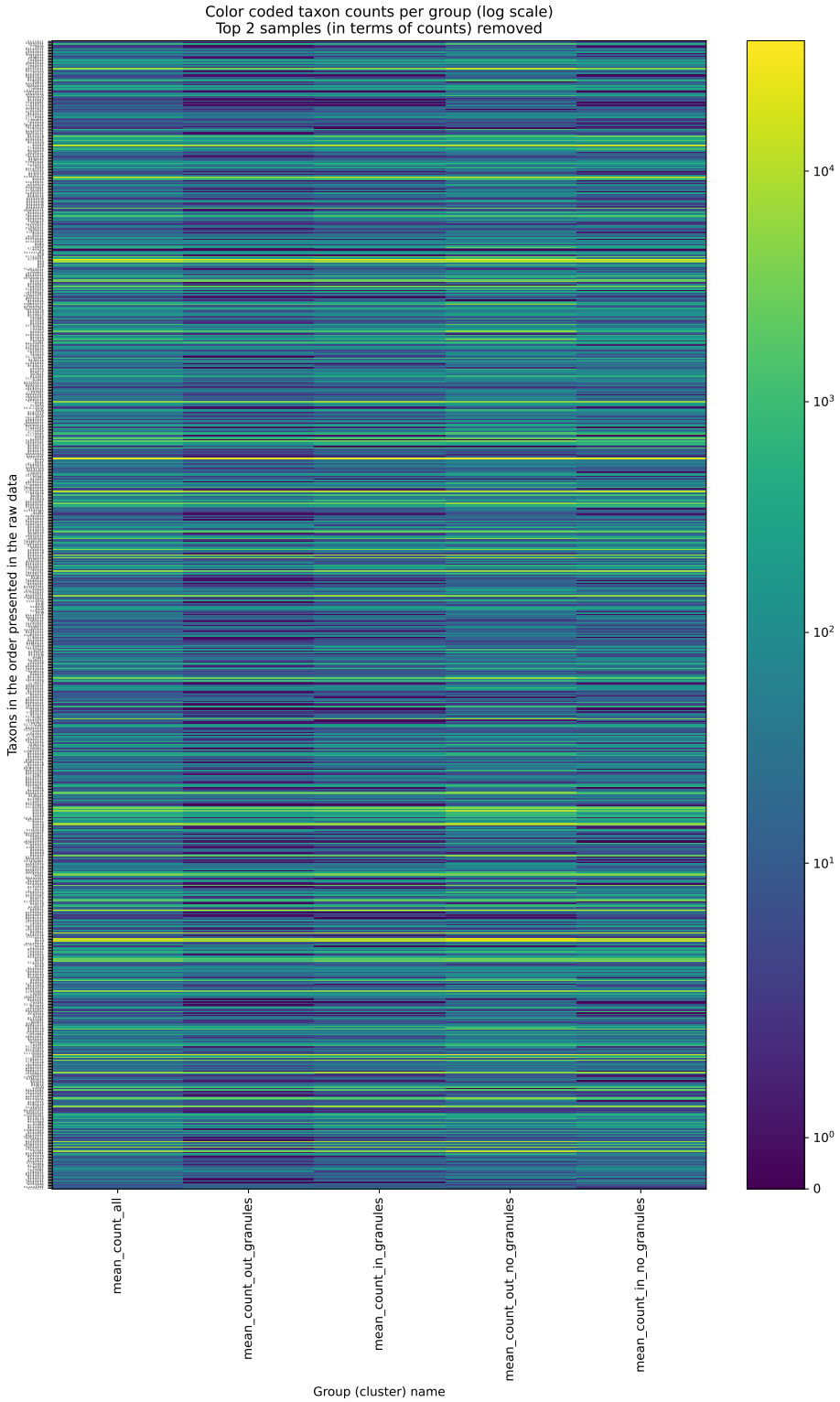


Figure 12: **Taxon counts per clustered group**

Conclusions on dimensionality reduction

The previous section described dimensionality reduction techniques that help to identify correlated variables and to extract usefull information. For example, by evaluating correlations, the PCA method reduced 851 taxons to 2 principal components that account for 83.4% of the variation in the data. The taxons that contributed most to the overall variance were identified by evaluating the PC loadings.

Inclusion of the response variable (granule formation) in the bi-plots of the principal components made it possible to identify clustered taxons. The cluster limits were defined using the standard deviation of the samples in the reduced PC1 and PC2 arrays. Useful information about the populations of taxons that formed granules was extracted by calculating the Bray-Curtis index and examining the taxon counts. The ranked differences between taxon counts inside and outside the cluster showed a large difference in *g Candidatus Accumulibacter* counts. Nevertheless, the low Bray-Curtis index shows that the use of taxon counts to assess wastewater treatment plants requires further optimisation in sampling and sample processing methods.

4.2 Truncated Singular Value Decomposition (SVD)

The first 2 principal components obtained using Trucated SVD are expected to explain a large part of the variation in the dataset. To compare the effect of the data distribution on clustering and correlation, the dimensionality reduction of the dataset was was evaluated with and without preprocessing.

Test method: Truncated SVD was performed on the following datasets:

- Raw data
- Standard scaled data
- MinMax scaled data

Figures 13, 14, and 15 show Bi-plots of the first 2 principal components. The percent of variance explained is shown on the x and y axis. The first 2 principal components explain 79.5% of the variance in the raw data. In contrast, the first 2 principal components explain respectively only 28.5% and 27.3% of the variance in the standard scaled and the MinMax scaled data.

In all 3 plots, a cluster (yellow dots) of samples with granules is visible. This demonstrates that any of the preprocessing methods can be used to separate granule forming microbial communities from communities that do not form granules.

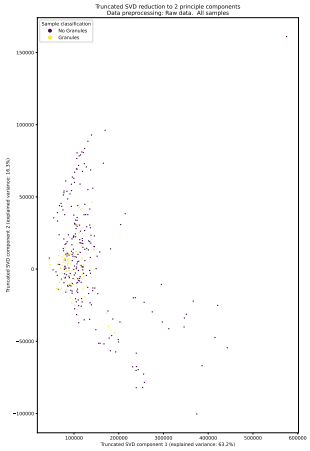


Figure 13: Raw data

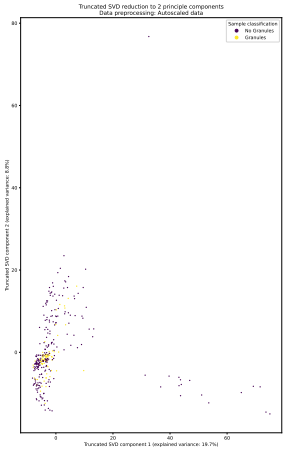


Figure 14: Standard scaled data

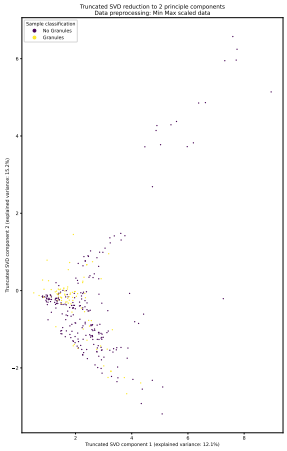


Figure 15: MinMax scaled data

Conclusion

The truncated SVD method is known to effectively reduce the dimensionallity of word count datasets. The dataset of taxon counts used in this study is formally similar to a word count data set. It was shown that the truncated SVD method effectively reduced the dimensionality of the raw data. Use of the PCA method as describe above reduced the dimensionality of the raw dataset containing all samples. The first 2 principal components obtained using PCA explain, respectively, 68.1 % and 15.3% of the data (total = 83.4%). In comparison, using the truncated SVD method the first 2 principal components explain respectively 63.2 % and 16.3% of the data (total = 79.5%). In contrast to the results using raw data, application of the truncated SVD to the dataset after preprocessing using the standard scaler or the MinMax scaler methods did not reduce the dimensionality as much. With raw, standard scaled or MinMax scaled data the Bi-plot of the first 2 principal components obtained using truncated SVD revealed a cluster of granule containing samples.

Machine learning

The previous sections described dimensionality reduction techniques that can be used to extract information from highly multivariate datasets. The following sections describe the use of machine learning methods to make predictive models. The term "Machine Learning" means that the data are split into 1) a training set that the model uses to learn the decision rules or predictor to response mapping and 2) a test set used to test the accuracy of predictions made from previously unseen data.

4.3 Logistic regression

The Logistic regression algorithm was implemented as a machine learning method. The default L2 (Ridge regression) regularization technique was used to introduce a penalty term intended to reduce overfitting. Logistic regression was performed on raw data.

The machine learning model was trained using 75% of the data. The model was tested on the remaining 25% of the previously unseen data. The highest accuracy was obtained after standard scaling pretreatment of the dataset. The accuracy score was 90%. This means that granule formation was correctly predicted from taxon counts in at least 90% of the experiments. These high accuracy scores indicate that logistic regression of standard scaled data can be used to predict granule formation and to identify the most important taxons.

Figure 16 shows the top ranked taxons in terms of the value of the coefficient of the logistic regression model when the mixed granule samples were removed. The genus *Dechloromonas* is the most important taxon in the logistic regression model. The negative value of the coefficient indicates that the presence of *Dechloromonas* is associated with failure to form AGS granules. *Dichloromonas* is known to be enriched in AGS flocs [2]. The presence of *Dichloromonas* might prevent the transition from flocs to granules.

The genus *Tahibacter* is associated positively with AGS granules. In low strenghth wastewater, *Tahibacter* were found to proliferate strongly during long-term maintenance of aerobic granules [17]. Also, the family *Xanthomonodaceae* is associated positively with AGS granules. In 4 litre sequencing batch reactors, the family *Xanthomonodaceae* with EPS secreting functions was enriched under 6-day SRT [18].

Data treatment	Accuracy (mixed granules/flocs removed)
Raw (no treatment)	0.83
Standard scaler	0.90
MinMax scaler	0.88

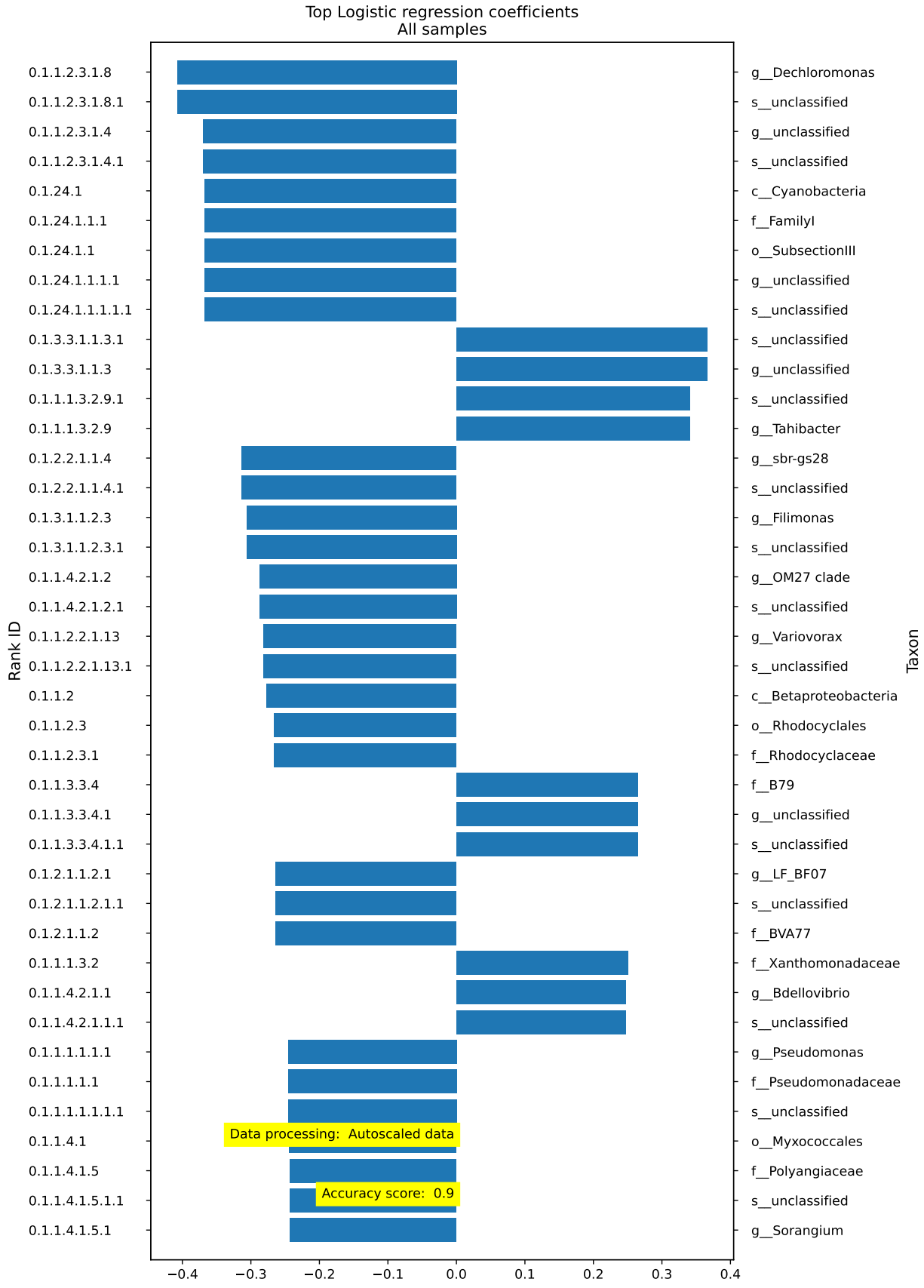


Figure 16: **Top ranked taxons by logistic regression coefficients**

4.4 Decision Tree

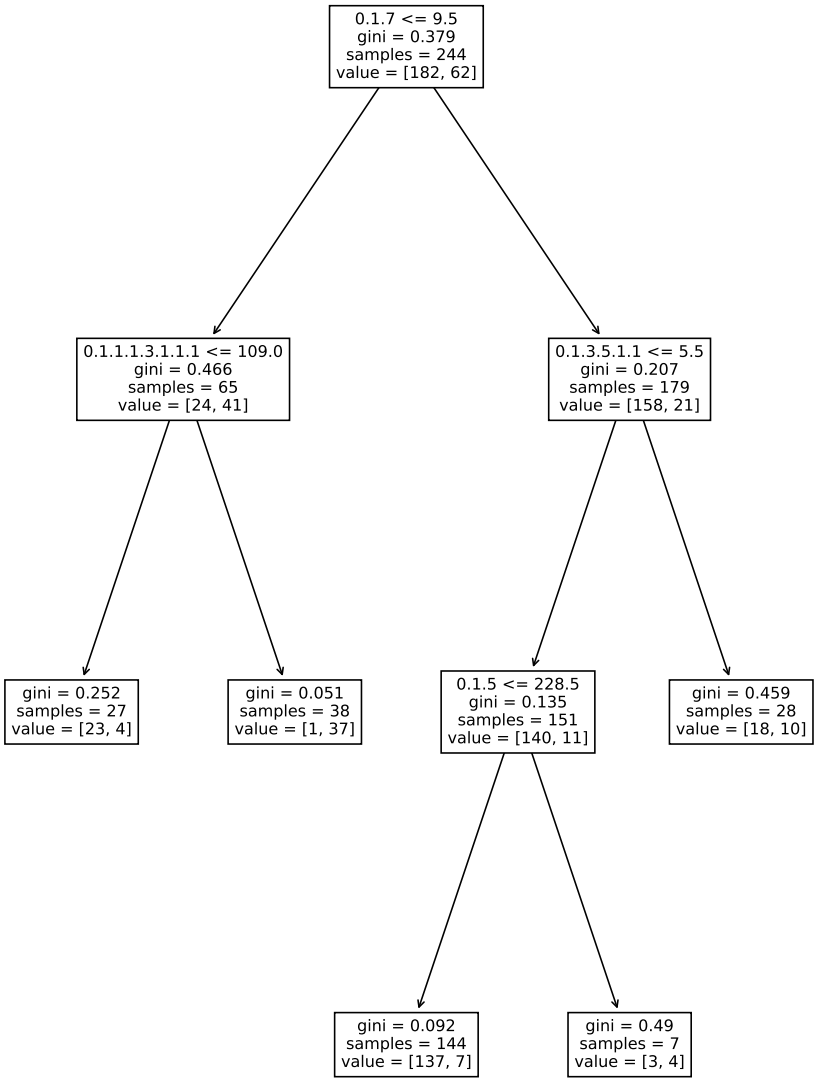
The purpose of the decision tree model is to predict granule formation from sample taxon counts. The tree defines a series of questions (branching nodes) that aim to yield a pure endpoint (leaf) where the samples are granules or mixed liquor. A pure leaf has a Gini index of 0. A leaf where half the samples are granules and half are mixed liquor has a Gini index of 0.5. The taxons identified in the branching nodes are critical to determining the classification to granules or to mixed liquor. The Decision tree algorithm was implemented as a supervized machine learning method.

The model creates a tree of branching nodes that present a question that results in sample classification into 2 bins (branches). Additionally the model reports the purity of the split achieved by answering the question, the number of pertinent samples concerned by the question, and the number of samples in each bin. If the answer to the question is "No", then the samples are classified in the left branch. If the answer is "Yes", then the samples are classified in the right branch.

Test method: The Decision tree model parameters are selected to optimize the accuracy, generalization, and interpretability of the tree. The maximum depth of the tree was set to limit the number of nodes in series to 3 in order to improve interpreatability and generalization. The minimum number of samples required to split a node was set to 50 in order to improve generalization. The default Gini criterion was used to measure the purity of the split at each node. The lower the value of Gini, the higher the purity of the split.

The machine learning model was trained using 75% of the data. The model was tested on the remaining 25% of the previously unseen data.

Decision tree to predict granule formation based on community composition (taxon counts).
Data: Results_aline_all_swarm. Data preprocessing: Raw data
Number of taxons to classify: 851. Number of samples: 326.
Predictors: counts per taxon. Response: Observation of granules.



Prediction accuracy: 0.84

Figure 17: **Decision tree (all samples)**

The prediction accuracy was 84% when all 326 samples were considered. Two almost pure leaves were identified in the tree shown in Figure 17. The first leaf (Gini = 0.051, left side) is reached under the following conditions:

1. *p_Verrucomicrobia* (0.1.7) count > 9.5
2. *s_unclassified* (0.1.1.1.3.1.1.1) count ≤ 109

Following this path results in 97% of the samples (1- 1/38) having granules.

The second leaf (Gini = 0.092, right side) is reached under the following conditions:

1. *p_Verrucomicrobia* (0.1.7) count ≤ 9.5
2. *f_unclassified* (0.1.3.5.1.1) count > 5.5
3. *c_unclassified*(0.1.5.1) count > 228.5

Following this path results in 95% of the samples (1 - 7/144) being classified as mixed liquor (not granules).

This result suggest that the presence of *p_Verruccomicrobia* has a postive impact on the production of granules. Since *p_Verruccomicrobia* is known to degrade complex polysacharides [14], its presence might help to produce the volitile fatty acids required by AGS producing microorganisms.

Figure 18 shows the counts per sample. Note: the samples are not in chronological order. The figure shows that the importance of a taxon determined by the decision tree is not the the same as the abundance of the taxon.

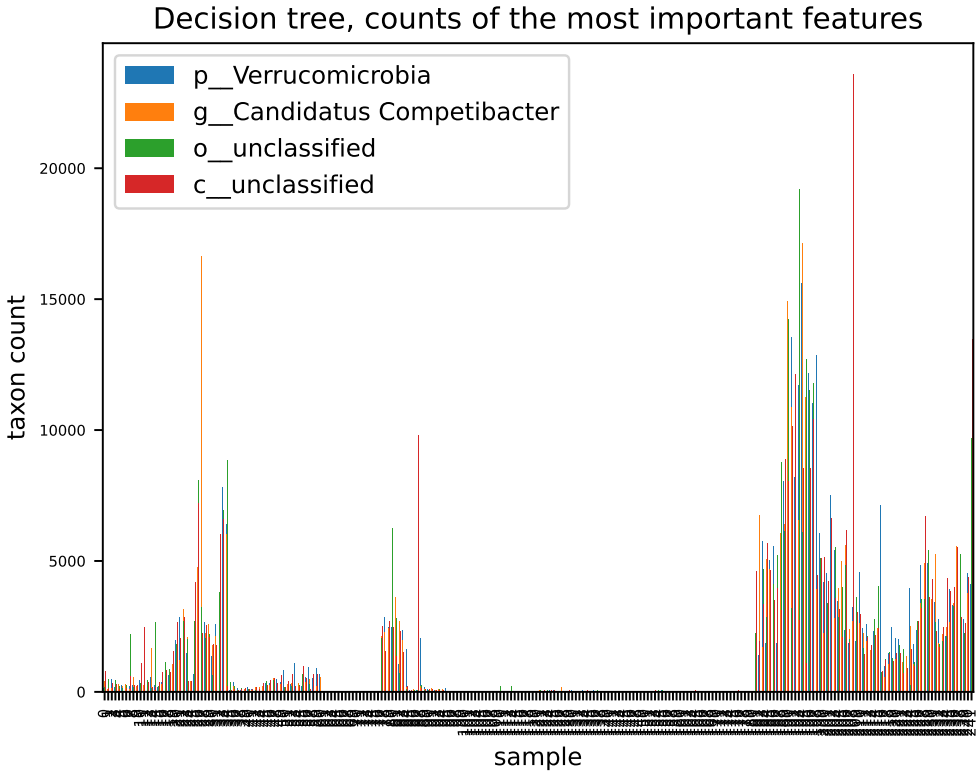


Figure 18: Decision tree, most important features

Summary of Decision Tree classifier metrics

Data treatment	Accuracy
Raw (no treatment)	0.87
Standard scaler	0.84
MinMax scaler	0.84

4.5 Random Forest

A Random forest is produced from replicates of Decision trees that are reinitialized before each run. Like the Decision tree model, the Random forest model parameters are selected to optimize the accuracy, generalization, and interpretability of the trees.

Test method: The Random forest was composed of 100 Decision trees parametrized like the Decision tree in the previous section. The Decision tree model parameters are selected to optimize the accuracy, generalization, and interpretability of the tree. The maximum depth of the tree was set to limit the number of nodes in series to 3 in order to improve interpreatability and generalization. The minimum number of samples required to split a node was set to 50 in order to improve generalization. The default Gini criterion was used to measure the purity of the split at each node. The lower the value of Gini, the higher the purity of the split.

Since the highest accuracy score of the Decision Tree classifier was obtained with raw data, raw data was also used to train the Random Forest classifier. The machine learning model was trained using 75% of the data. The model was tested on the remaining 25% of the previously unseen data.

The relative importance of the taxons was calculated by first ranking the relative feature (taxon) importances of the fitted model and then taking the sum of the counts of each feature. The classifier calculates the feature importance from the purity of the branches produced by the node.

Figures 19 shows the most important taxons in the random forest.

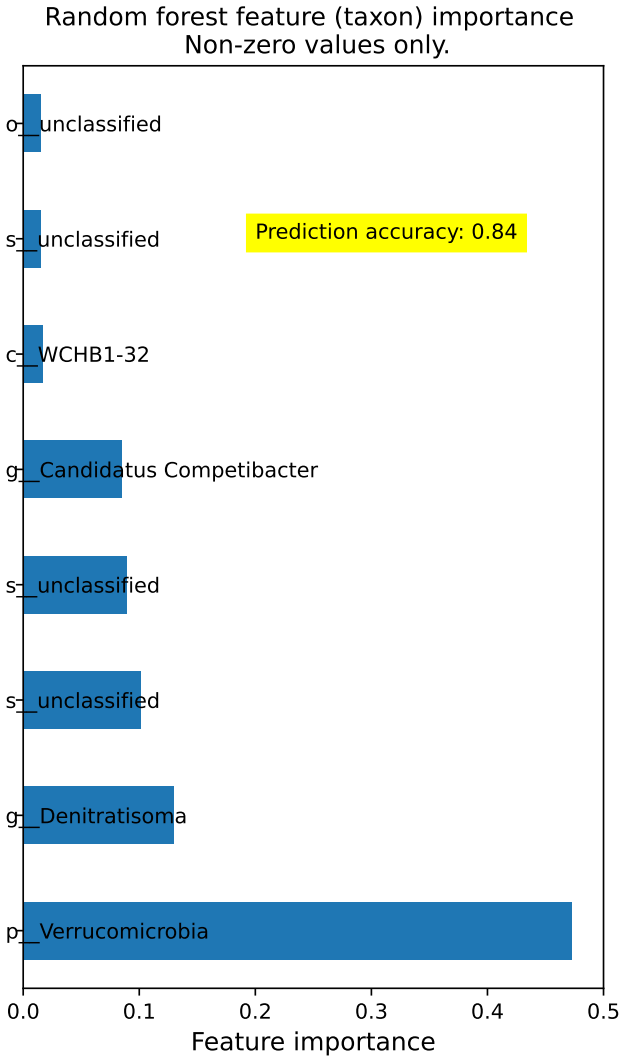


Figure 19: Random forest, most important features

Figure 20 shows the counts of most important taxons in the random forest trained with respectively, all samples, and with mixed granule samples removed. As shown in Figure 20, when all samples are considered *p verrucomicrobia* and *Denitratasoma* are the most important features in the predictive model. However, the count of the fifth most important taxon, *g Candidatus Competibacter*, is more than 30X higher than the count of *p verrucomicrobia*. This result shows that definition of the community composition associated with granule formation requires more than the identity of the taxons and the taxon counts. The definition also requires an importance ranking that does not depend on the counts. In the case of the random forest algorithm, taxon importance and definition of the community composition and critical taxons depends on a series of questions about the relative counts of critical taxons.

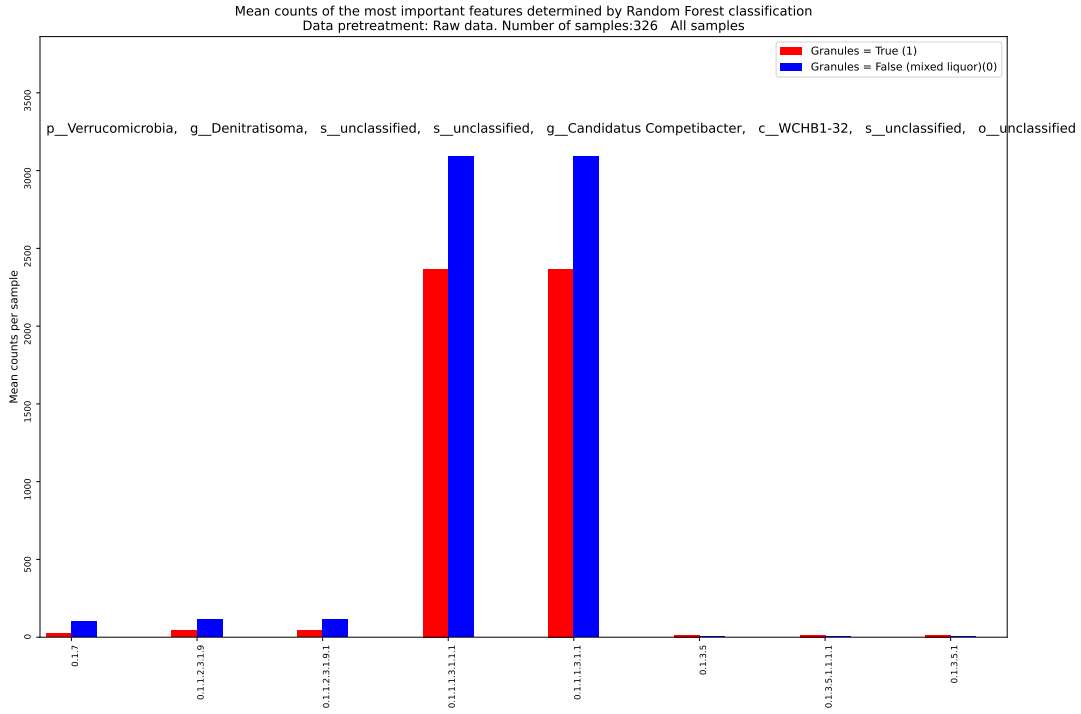


Figure 20: **Random forest, counts of the most important features**

A confusion matrix is used to evaluate the prediction accuracy in terms of True positives, True negatives, False positives, and False negatives where True = granule formation and False = no granule formation. Figure 21 shows the confusion matrices for random forests trained with with, respectively, all samples, and with mixed granule samples removed.

The confusion matrix shows the results of classification of the test samples (25% of the data) reserved for testing the trained Random forest predictive model. Classification metrics can be obtained from the confusion matrix.

Figure 21 shows the results of the model built using all samples. The accuracy is the number of correct predictions of True positives (58) and True negatives (11) divided by the total number of samples (82). The accuracy was 84% ((58 + 11)/82 = 0.84). The model precision is the ratio of correctly predicted True positives (58) to the number of samples predicted as positive (58 + 11). The precision was 84% (58/(58+11) = 0.84). The model recall is the ratio of correctly predicted True positives (58) to the actual number of True positives (58 + 2). The model recall was 97% (58/(58+2) = 0.97). The F1 score is the harmonic mean of precision and recall. The model F1 score was 0.9 ((2 x 0.84 x 0.97)/(0.84 + 0.97) = 0.9).

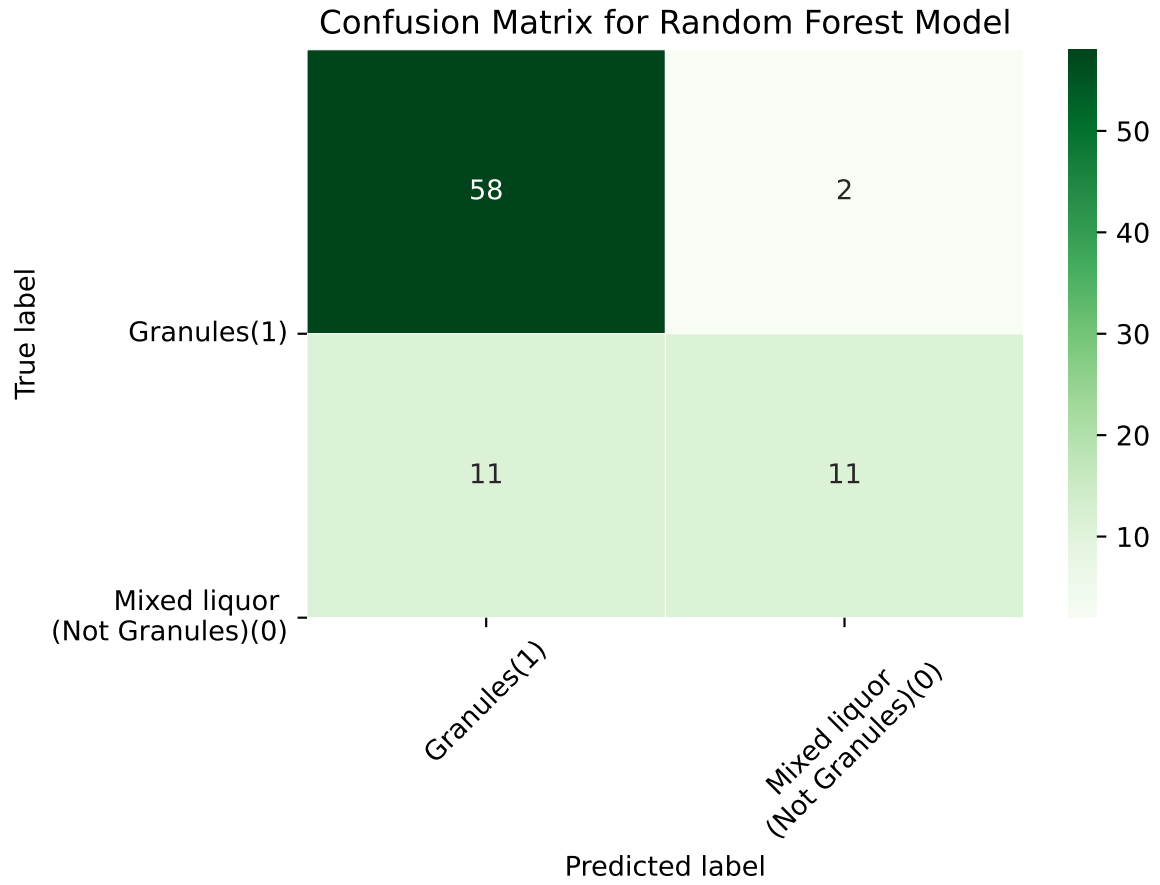


Figure 21: All samples

Summary of Random forest classifier metrics

Model type	result
Accuracy	0.84
Precision	0.84
Recall	0.97
F1 score	0.90

4.6 Naive Bayes Classifier

The Multinomial Naive Bayes classifier less accurately predicted granule formation than did the other models.

Test method: The smoothing parameters that account for feaures not present in the learning samples were set at their default values. The machine learning model was trained using 75% of the data. The model was tested on the remaining 25% of the previously unseen data. A confusion matrix is used to evaluate the prediction accuracy in terms of True positives, True negatives, False positives, and False negatives where True = granule formation and False = no granule formation.

The highest prediction accuracy of 0.77 was achieved when the data was MinMax scaled prior to training the model. Date pretreatment by standard scaling was not evaluated because the classifier does not accept negative values.

As shown in Figure 22, the model predicted many false positives and false negatives.

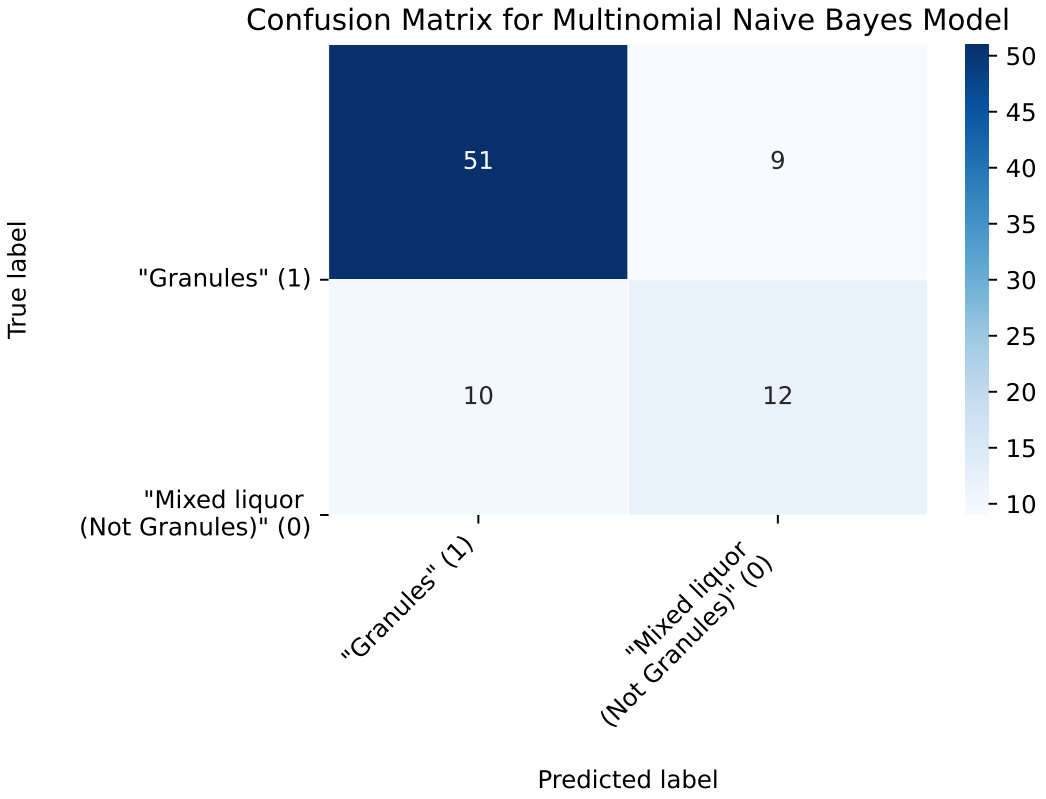


Figure 22: All samples

Summary of Multinomial Naive Bayes classifier metrics

Data treatment	Accuracy
Raw (no treatment)	0.67
Standard scaler	Not evaluated
MinMax scaler	0.77

Comparison of machine learning model prediction accuracy

Following the conventional machine learning practice, the data were split into training (75% of the data) and test sets (25% of the data). The models were trained and the prediction accuracy was assessed using the previously unseen test data set. The highest prediction accuracy (90%) was attained using the Logistic regression model with standard scaled data.In general, the accuracy score is used to compare machine learning models. A high accuracy score implies that the model makes a robust mapping of the predictor data set to the response dataset.

Model type	Accuracy
Logistic regression	0.90
Decision tree	0.87
Random forest	0.84
Multinomial Naive Bayes	0.77

5 Conclusions

The main conclusions of this study are:

- In terms of taxon counts, the populations found in granules are not so different from the populations found in the mixed liquor (Bray-Curtis index = 0.14).
- The taxons *c Betaproteobacteria*, *f Rhodocyclales*, *o Rhodocyclales*, *p Saccharibacteria*, and *g Candidatus Accumulibacter* are more abundant inside the granule cluster limits than outside the cluster limits.
- Using raw taxon counts to distinguish sample fractions, the logistic regression function most accurately predicts granule fractions as distinguished from mixed liquor (Prediction accuracy: 90%).
- The taxon *Verrumicrobia* is an important determinant of granules (revealed by the Decision tree and the Random forest) and might influence granule formation.
- The presence of *Dechloromonas* adversely affects sample classification as granules.

Subjects for future study include:

- The functional differences between the populations that form granules and the populations that form mixed granules/flocs and the populations that do not form granules (taxon distance evaluated using the Bray-Curtis index).
- Implementation of the logistic regression function as a machine learning model to predict granule formation in larger datasets obtained from long-term sampling of wastewater treatment plants.
- The functions of *Verrumicrobia* and its role in degradation of complex polysacharides.
- Application of Random forest and Logistic regression machine learning models to long-term studies of large datasets obtained from wastewater treatment plants to identify taxons that are robust indicators of granule formation and of failure to form granules.
- The effect of data preprocessing and normalization on model prediction accuracy.

This study demonstrated the use of machine learning techniques to make predictive models of highly multivariate data that had non-normal distributions. In this study, machine learning methods are appropriate because the reliance on the prediction accuracy score using previously unseen test data validates the data preprocessing methods and the model. Consequently, a machine learning model that accounts for very complex interactions between prediction variables can be developed.

The methods used to make this study describe neither the chemical mechanisms, the metaboloc pathway, nor the microbial ecology of AGS formation. However, the methods do reveal predictable relationships between taxon counts and AGS granules and AGS mixed liquor. Consequently, the results suggest subjects for further investigation of chemical mechanisms, the metaboloc pathway, and microbial ecology.

Additionally, the methods might be useful for the characterisation of full-scale wastewater treatment plants where the objective is to optimise performance rather than to understand fundamental mechanisms. For example, the methods that were applied during this study might be useful for determining the AGS formation potential of samples from full-scale wastewater treatment plants. Methods to compare the taxonomic profiles of AGS wastewater treatment systems to reference taxonomic profiles from high performing AGS systems might be useful for trouble shooting and optimization.

Finally, the results obtained during this study are dependant on the dataset. The data was obtained during long term experiments conducted according to a plan with different objectives than those of this study. If machine learning methods were applied in a future study, then the experimental design should aim to generate a dataset that covers a wide range of potential operating setpoints and conditions. In the case of a study conducted at a full-scale wastewater treatment plant, the dataset should include the full range of real operating conditions.

6 Acknowledgements

Thank you to Aline Adler for all of the work to generate and analyse the data, to Christof Holliger for making it possible to do the project in his lab, and to Laëtitia Cardona for supplying the background information and explanations that I needed to understand the context and how the data was obtained.

My personal objective in this study was to improve my data science skills and my knowledge in the field of environmental biotechnology. To reach this objective, I asked Professor Christof Holliger, head of the LBE (Laboratoire de biotechnologie environnementale) at the EPFL if he had a suitable data set for me to work on. He gave me access to a large data set of bioreactor experiments conducted with the aim of understanding the metagenomics and the metabolic pathways of AGS. My work does not aim to understand metagenomics and metabolism. My work was to apply different methods with the aim of identifying methods to obtain accurate predictions of granule formation and to identify the critera (predictors) from laboratory experiments that are important to granule formation.

References

[1] Aline Sondra Adler. "The effect of different organic substarates on the microbial communities of aerobic granular wastewater treatment sludge". PhD thesis. 2019.

[2] Aline Adler and Christof Holliger. "Multistability and Reversibility of Aerobic Granular Sludge Microbial Communities Upon Changes From Simple to Complex Synthetic Wastewater and Back". In: *Frontiers in Microbiology* 11 (2020). DOI: 10.3389/fmicb.2020.574361.

[3] Amoeba. *Relationship between SVD and PCA. How to use SVD to perform PCA?* Accessed on 04.08.2023. URL: https://stats.stackexchange.com/questions/134282/relationship-between-svd-and-pca-how-to-use-svd-to-perform-pca.html.

[4] Kevin Dunn. *Process Improvement Using Data*. Accessed on 04.08.2023. URL: https://learnche.org/pid/preface/index.html.

[5] Tao Guo et al. "Aerobic granular sludge coupling with Fe–C in a continuous-flow system treating dyeing wastewater on-site". In: *Environmental Technology and Innovation* 30 (2023). DOI: 10.1016/j.eti.2023.103065.

[6] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[7] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[8] Scikit-learn. *Decision tree classifier*. Accessed on 07.08.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[9] Scikit-learn. *Dimensionality reduction using truncated SVD*. Accessed on 04.08.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html#sklearn.decomposition.TruncatedSVD.html.

[10] Scikit-learn. *Logistic Regression (aka logit, MaxEnt) classifier*. Accessed on 04.08.2023. URL: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.html.

[11] Scikit-learn. *Multinomial Naive Bayes classifier*. Accessed on 07.08.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB.html.

[12] Scikit-learn. *Random forest classifier*. Accessed on 07.08.2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.html.

[13] Scipy. *Compute the Bray-Curtis distance*. Accessed on 04.08.2023. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.braycurtis.html.

[14] A Sichert, C.H. Corzett, and M.S. Schechter. "Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan". In: *Nat Microbiol* 5 (2020), pp. 1026–1039. DOI: 10.1038/s41564-020-0720-2.

[15] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[16] DG Weissbrodt et al. "Bacterial Selection during the Formation of Early-Stage Aerobic Granules in Wastewater Treatment Systems Operated Under Wash-Out Dynamics". In: *Front Microbiol* 3 (2012). DOI: 10.3389/fmicb.2012.00332.

[17] H Zhang et al. "Reproducibility of Aerobic Granules in Treating Low-Strength and Low-C/N-Ratio Wastewater and Associated Microbial Community Structure." In: *Processes* 10 (2022). DOI: 10.3390/pr10030444.

[18] Zhiming Zhang et al. "Understanding of aerobic sludge granulation enhanced by sludge retention time in the aspect of quorum sensing". In: *Bioresource Technology* 272 (2019), pp. 226–234. DOI: 10.1016/j.biortech.2018.10.027.