PhD School of Pharmaceutical Sciences XXXII cycle
"Sapienza" University of Rome

Machine Learning Applications to Essential Oils and Natural Extracts

Doctoral Dissertation
Submitted by

Alexandros Patsilinakos

Department of Chemistry and Technologies of Drug
Faculty of Pharmacy and Medicine
"Sapienza" University of Rome

Supervisor

Professor **Rino Ragno**

Department of Chemistry and Technologies of Drug, “Sapienza” University of Rome, P.le Aldo Moro, 5 - 00185 - Rome, Italy.

LIST OF PUBLICATIONS:

Rino Ragno, Rosanna Papa, Alexandros Patsilinakos, Gianluca Vrenna, Stefania Garzoli, Vanessa Tuccio, ErsiliaVita Fiscarelli, Laura Selan, and Marco Artini. "Essential oils against bacterial isolates from cystic fibrosis patients by means of antimicrobial and unsupervised machine learning approaches." *Scientific Reports* 10, no. 1 (2020): 1-11.

Garzoli, S., L. Antonini, A. Troiani, C. Salvitti, P. Giacomello, A. Patsilinakos, R. Ragno, and F. Pepi. "Gas-phase structures and thermochemical properties of protonated 5-HMF isomers." *International Journal of Mass Spectrometry* 447 (2020): 116237.

Marrocco, Fabio Altieri, Elisabetta Rubini, Giuliano Paglia, Silvia Chichiarelli, Flavia Giamogante, Alberto Macone, Giacomo Perugia, Fabio Massimo Magliocca, Aymone Gurtner, Bruno Maras, Rino Ragno, Alexandros Patsilinakos, Roberto Manganaro, Margherita Eufemi. "Shmt2: A Stat3 Signaling New Player in Prostate Cancer Energy Metabolism." *Cells* 8, no. 9 (2019): 1048.

Alexandros Patsilinakos, Marco Artini, Rosanna Papa, Manuela Sabatino, Mijat Božović, Stefania Garzoli, Gianluca Vrenna, Raissa Buzzi, Stefano Manfredini, Laura Selan, Rino Ragno. "Machine Learning Analyses on Data including Essential Oil Chemical Composition and In Vitro Experimental Antibiofilm Activities against Staphylococcus Species." *Molecules* 24, no. 5 (2019): 890.

Giulia Stazi, Cecilia Battistelli, Valentina Piano, Roberta Mazzone, Biagina Marrocco, Sara Marchese, Sharon M Louie, Clemens Zwergel, Lorenzo Antonini, Alexandros Patsilinakos, Rino Ragno, Monica Viviano, Gianluca Sbardella, Alessia Ciogli, Giancarlo Fabrizi, Roberto Cirilli, Raffaele Strippoli, Alessandra Marchetti, Marco Tripodi, Daniel K Nomura, Andrea Mattevi, Antonello Mai, Sergio Valente. "Development of alkyl glycerone phosphate synthase inhibitors: Structure-activity relationship and effects on ether lipids and epithelialmesenchymal transition in cancer cells.” *European journal of medicinal chemistry* 163 (2019): 722-735.

Maxim B Nawrozkij, Mariantonietta Forgione, Alexandre S Yablokov, Alessia Lucidi, Daniela Tomaselli, Alexandros Patsilinakos, Cristina Panella, Gebremedhin S Hailu, Ivan A Kirillov, Roger Badia, Eva Riveira-Muñoz, Emmanuele Crespan, Jorge I Armijos Rivera, Roberto Cirilli, Rino Ragno, José A Esté, Giovanni Maga, Antonello Mai, Dante Rotili. “Effect of α-Methoxy Substitution on the Anti-HIV Activity of Dihydropyrimidin-4 (3 H)-ones.” *Journal of medicinal chemistry* 62, no. 2 (2018): 604-621.

Alexandros Patsilinakos, Rino Ragno, Simone Carradori, Stefania Petralito, Stefania Cesa. “Carotenoid content of Goji berries: CIELAB, HPLC-DAD analyses and quantitative correlation.” *Food chemistry* 268 (2018): 49-56.

L Antonini, S Garzoli, A Ricci, A Troiani, C Salvitti, P Giacomello, R Ragno, A Patsilinakos, B Di Rienzo, F Pepi. “Ab-initio and experimental study of pentose sugar dehydration mechanism in the gas phase.” *Carbohydrate research* 458 (2018): 19-28.

Manuela Sabatino, Dante Rotili, Alexandros Patsilinakos, Mariantonietta Forgione, Daniela Tomaselli, Fréderic Alby, Paola B Arimondo, Antonello Mai, Rino Ragno. “Disruptor of telomeric silencing 1-like (DOT1L): disclosing a new class of non-nucleoside inhibitors by means of ligand-based and structure-based approaches.” *Journal of computer-aided molecular design* 32, no. 3 (2018): 435-458.

Marco Artini, Alexandros Patsilinakos, Rosanna Papa, Mijat Božović, Manuela Sabatino, Stefania Garzoli, Gianluca Vrenna, Marco Tilotta, Federico Pepi, Rino Ragno, Laura Selan. “Antimicrobial and Antibiofilm Activity and Machine Learning Classification Analysis of Essential Oils from Different Mediterranean Plants against Pseudomonas aeruginosa.” *Molecules* 23, no. 2 (2018): 482.

Simone Carradori, Bruna Bizzarri, Melissa D'Ascenzio, Celeste De Monte, Rossella Grande, Daniela Rivanera, Alessanda Zicari, Emanuela Mari, Manuela Sabatino, Alexandros Patsilinakos, Rino Ragno, Daniela Secci. "Synthesis, biological evaluation and quantitative structure-active relationships of 1, 3-thiazolidin-4-one derivatives. A promising chemical scaffold endowed with high antifungal potency and low cytotoxicity." *European Journal of Medicinal Chemistry* 140 (2017): 274-292.

Clemens Zwergel, Brigitte Czepukojc, Emilie Evain-Bana, Zhanjie Xu, Giulia Stazi, Mattia Mori, Alexandros Patsilinakos, Antonello Mai, Bruno Botta, Rino Ragno, Denise Bagrel, Gilbert Kirsch, Peter Meiser, Claus Jacob, Mathias Montenarh, Sergio Valente. "Novel coumarin-and quinolinone-based polycycles as cell division cycle 25-A and-C phosphatases inhibitors induce proliferation arrest and apoptosis in cancer cells." *European Journal of Medicinal Chemistry* 134 (2017): 316-333.

Milan Mladenovic, Alexandros Patsilinakos, Adele Pirolli, Manuela Sabatino, Rino Ragno. "Understanding the Molecular Determinant of Reversible Human Monoamine Oxidase B Inhibitors Containing 2*H*-Chromen-2-One Core: Structure-Based and Ligand-Based Derived Three-Dimensional Quantitative Structure–Activity Relationships Predictive Models." Journal of Chemical Information and Modeling, 57(4), 787-814.

Nikolaos Papastavrou, Maria Chatzopoulou, Jana Ballekova, Mario Cappiello, Roberta Moschini, Francesco Balestri, Alexandros Patsilinakos, Rino Ragno, Milan Stefek, Ioannis Nicolaou. Enhancing activity and selectivity in a series of pyrrol-1-yl-1-hydroxypyrazole-based aldose reductase inhibitors: The case of trifluoroacetylation. *European Journal of Medicinal Chemistry* 130 (2017): 328-335.

Adriano Mollica, Sveva Pelliccia, Valeria Famiglini, Azzurra Stefanucci, Giorgia Macedonio, Annalisa Chiavaroli, Giustino Orlando, Luigi Brunetti, Claudio Ferrante, Stefano Pieretti, Ettore Novellino, Sandor Benyhe, Ferenc Zador, Anna Erdei, Edina Szucs, Reza Samavati, Szalbolch

Dvrorasko, Csaba Tomboly, Rino Ragno, Alexandros Patsilinakos, Romano Silvestri. "Exploring the first Rimonabant analog-opioid peptide hybrid compound, as bivalent ligand for CB1 and opioid receptors. Journal of enzyme inhibition and medicinal chemistry", *Journal of enzyme inhibition and medicinal chemistry* 32, no. 1 (2017): 444-451.

Acknowledgements

I would like to acknowledge everyone who played a role in my academic accomplishment.

I would like to express my sincere gratitude to my advisor and mentor Prof. Rino Ragno and his research group, where I had the opportunity to work in the last years.

Many thanks also to all the research groups with whom I worked in the realization of these projects, and in particular:

> Simone Carradori, Stefania Petralito, and Stefania Cesa for their contribution to the “Carotenoid content of Goji berries: CIELAB, HPLC-DAD analyses and quantitative correlation” paper.
>
> Marco Artini, Rosanna Papa, Mijat Božović, Stefania Garzoli, Gianluca Vrenna, Marco Tilotta, Federico Pepi, and Laura Selan for designing and performing the analytical chemistry, extraction, and biological experiments on Essential Oils.

Besides my advisor, I would like to thank the coordinator of the PhD School of Pharmaceutical Sciences Prof. Claudio Villani, for his insightful comments and encouragement.

I thank my fellow lab mates, Lorenzo Antonini and Manuela Sabatino, for the stimulating discussions, for the long days we were working together before deadlines, and for all the fun we have had in the last years. Also, I thank my friends Eleni Pontiki, Claudia Trivisani, Usha Singh and Roberto Manganaro.

Last but not the least, I would like to thank my family for supporting me throughout writing this thesis and my life in general.

Abstract

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that allow computers to learn without being explicitly programmed. Various are the applications of ML in pharmaceutical sciences, especially for the prediction of chemical bioactivity and physical properties, becoming an integral component of the drug discovery process. ML is characterized by three learning paradigms that differ in the type of task or problem that an algorithm is intended to solve: supervised, unsupervised, and reinforcement learning. In chapter 2, supervised learning methods were applied to extracts of *Lycium barbarum* L. fruits for the development of a QSPR model to predict zeaxanthin and carotenoids content based on routinely colorimetric analyses performed on homogenized samples, developing a useful tool that could be used in the food industry. In chapters 3 and 4, ML was applied to the chemical composition of essential oils and correlated to the experimentally determined associated biofilm modulation influence that was either positive or negative. In these two studies, it was demonstrated that biofilm growth is influenced by the presence of essential oils extracted from different plants harvested in different seasons. ML classification techniques were used to develop a Quantitative Activity-Composition Relationship (QCAR) to discover the chemical components mainly responsible for the anti-biofilm activity. The derived models demonstrated that machine learning is a valuable tool to investigate complex chemical mixtures, enabling scientists to understand each component's contribution to the activity. Therefore, these classification models can describe and predict the activity of chemical mixtures and guide the composition of artificial essential oils with desired biological activity. In chapter 5, unsupervised learning models were developed and applied to clinical strains of bacteria that cause cystic fibrosis. The most severe infections reoccurring in cystic fibrosis are due to *S. aureus* and *P. aeruginosa*. Intensive use of antimicrobial drugs to fight lung infections leads to the development of antibiotic-resistant bacterial strains. New antimicrobial compounds should be identified to overcome antibiotic resistance in patients. Sixty-one essential oils were studied against a panel of 40 clinical strains of *S. aureus* and *P. aeruginosa* isolated from cystic fibrosis patients, and unsupervised machine learning algorithms were applied to pick-up a small number of representative strains (clusters of strains) among the panel of 40. Thus, rapidly identifying three essential oils that strongly inhibit antibiotic-resistant bacterial growth.

Table of Contents

1 Introduction

1.1 What is Machine Learning?

Machine Learning (ML) is a branch of Artificial Intelligence (AI). Its foundation lies in algebra, statistics, probability, and it collects methods developed in the last decades of the twentieth century in various scientific disciplines which uses mathematical optimization and statistical methods to improve the performance of an algorithm in identifying patterns in data: computational statistics, information theory, pattern recognition, Bayesian methods, neuroscience, artificial neural networks, dynamical systems theory, image processing, data mining, adaptive algorithms, and last but not least bioinformatics and cheminformatics. In the field of computer science, machine learning is a variant of traditional programming in which a machine develops the ability to learn something from data independently, without explicit programmed. Arthur Samuel coined the term Machine Learning in 1959, which describes it as "the field of study that gives computers the ability to learn without being explicitly programmed"[1]. Arthur Samuel identifies two distinct approaches. The first method, referred to as a neural network, develops general-purpose machine learning machines that learn from a randomly connected switching network, following a reward-and-punishment-based learning routine (reinforcement learning). The second, more specific method is to reproduce the equivalent of a highly organized network designed to learn only specific activities. The second procedure, which requires supervision and requires reprogramming for each new application, is much more computationally efficient.

A more formal definition was provided by Tom Mitchell which states: "A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**."[2]. For example, if we considering playing checkers, the experience of playing many games of checkers is **E**, the task of playing checkers is **T**, and the probability that the program will win the next game is **P**. A machine learning algorithm build a mathematical model by training itself on sample data (training data). This model can be used to make predictions and decisions on new, unseen samples (unknown data).

The generic workflow of the development of an ML model consists of six steps, independent of the algorithm adopted[3]:

1. **Collect and prepare the data** in a format that can be given as input to the algorithm. The data needs to be cleaned and pre-processed in a structured format. The accuracy of the learned function depends on the input object representation. The input object is transformed into a feature vector, which contains several features that are descriptive of the object. The number of features should contain enough information to predict the output accurately.
2. Perform a **feature selection** of the most relevant properties to the learning process. Some data features need to be removed, and a subset of the most important features needs to be obtained.
3. **Chose the ML algorithm** that mostly suits a specific class problem. Selecting the best ML algorithm and therefore determining the structure of the learned function is critical for getting the best results.
4. **Select the model's parameters**. Some learning algorithms require the user to determine specific control parameters. These parameters are regulated by optimizing performance on a subset of the training set (validation set) or via cross-validation.
5. **Train** of the algorithm using a part of the dataset as training data.
6. **Evaluate the model performance**. The model must be tested against unseen data to be validated against various performance parameters.

The overall process in unsupervised learning modeling can be summarized in **Figure 1**.

Three broad machine learning paradigms exist that differ on the type of task or problem an algorithm is intended to solve, how it is being trained, and the input/output type[4].

- **Supervised Learning**: the algorithm is given examples in the form of possible inputs and their respective desired outputs, and the goal is to extract a general rule that associates the input with the correct output.
- **Unsupervised Learning**: the algorithm aims to find a structure in the inputs provided, without the inputs being labeled by any means.
- **Reinforcement Learning:** the algorithm learns from a series of reinforcements (rewards and punishments) by interacting with a dynamic environment in which it

tries to reach a goal by making decisions to achieve the highest reward (for example, to design new molecules with specific desired properties[5]). It learns through trial and error, and a sequence of successful decisions will result in the process being reinforced because it best solves the problem in question.

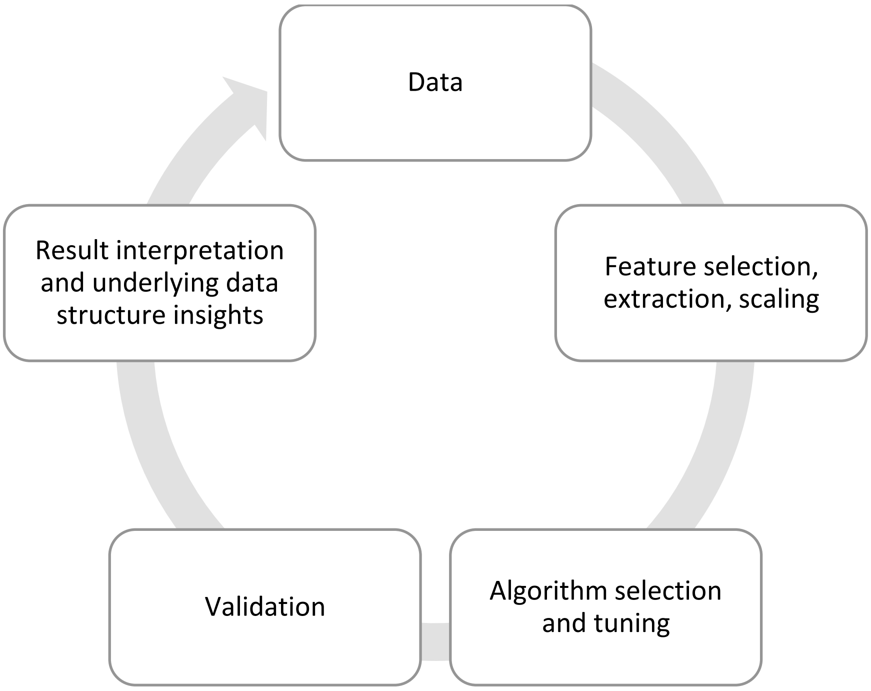

Data
Feature selection, extraction, scaling
Algorithm selection and tuning
Validation
Result interpretation and underlying data structure insights

Figure 1. The machine learning analysis process

1.2 Supervised Learning

In supervised learning, the algorithm task is that of learning a mathematical function that maps an input (typically a vector **x**) to an output variable (y) by observing input-output pairs (training set). The inferred function can be used for mapping new examples.

Formally, given a **training set** of *N* example input-output pairs

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots (x_N, y_N)$$

where each pair was generated by an unknown function $y = f(x)$, discover a **hypothesis** function *h*, drawn from a **hypothesis space** *H* of possible functions that approximate the true function *f*. A second sample of (x_i, y_i) pairs called **test set** is needed to evaluate the quality of the hypothesis and, therefore, of predictions[4]. The hypothesis *h* generalizes well if it accurately predicts the outputs of the test set.

Supervised learning problems are categorized into classification and regression problems.

In a classification problem, samples belong to two or more classes, and the response values are categorical. For example, we want to predict if a molecule is active or inactive versus a biological target. Therefore, we are trying to map input variables to a discrete function. Logistic Regression, *k*-Nearest Neighbors (*k*NN)[6], Support Vector Machines (SVM) [7], Random Forests[8], Gradient Boosting Machines[9], and Neural Networks are methods commonly used in classification.

In a regression problem, the output value to predict consists of one or more continuous variables, meaning that we are trying to map input variables to a continuous function. For example, predicting the molecule's pIC50 (that inhibits an enzyme) as a function of its physicochemical properties is a regression problem. Linear regression, Nonlinear regression, k-Nearest Neighbors (*k*NN)[10], Support Vector Machines (SVM) [7], Random Forests[8], Gradient Boosting Machines[9], and Neural Networks are methods commonly used in regressions.

To develop machine learning models with good predictive capabilities, it is necessary to adopt specific methods that allow establishing:

- which are the best settings of the specific parameters of the algorithms taken into consideration for the development of the models (hyperparameter fine-tuning);
- the predictive capabilities and intrinsic robustness of the model, using appropriate evaluation metrics.

The confidence that the trained model will generalize well on the unseen data can never be high without proper model validation. Fitting the parameters of a prediction function and testing it on the same data is a methodological mistake. A model that would reproduce the samples' labels that it has just seen would have an excellent score but would fail to predict anything valuable on yet-unseen data (overfitting). Model validation helps assure that the model performs well on new data and, therefore, it is robust. It is possible to select the best model, the parameters, and the accuracy metrics through model validation and grid search techniques. Once the models are trained, it is possible to test their reliability through two important validation procedures, such as cross-validation (CV, or internal validation, carried out taking into consideration the same dataset used in the development of the model) and external validation (applied when new data are used). When evaluating different settings (hyperparameters) for machine learning estimators, there is a risk of overfitting on the test set because the parameters can be tweaked until the estimator performs optimally. Knowledge of the test set can leak into the model, and evaluation metrics no longer report on generalization performance. A solution to this problem is to hold out yet another part of the dataset (validation set). Training progresses on the training set, after which evaluation is done on the validation set, and when the learning process seems to be successful, the final evaluation can be done on the test set. By partitioning the available data into three sets, we drastically reduce the number of samples that can be used to learn the model, and the validation outcomes can depend on a distinct random choice for the pair of train and validation sets. The cross-validation (CV) procedure is a solution to this problem. The validation set is no longer needed when doing CV, as it is intrinsically created at each iteration. In the basic approach, called *k*-fold CV, the training set is split into *k* smaller sets (**Figure 2**).

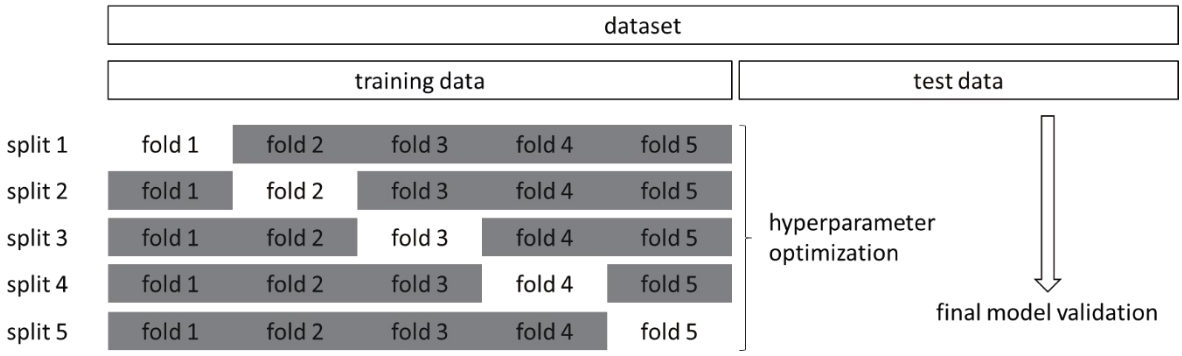


Figure 2. Subdivision of the dataset in the *k*-fold cross-validation with *k* = 5.

For each of the *k* folds, a model is trained using $k - 1$ of the folds as training data, and the resulting model is validated on the remaining part of the data by using a performance measure. The performance measure reported by *k*-fold cross-validation is the average of the values computed in the loop. This approach has a significant advantage in problems such as inverse inference, where the number of samples is small, as it does not waste too much data (as is when fixing an arbitrary validation set), although it can be computationally expensive.

A particular case of *k*-fold cross-validation is represented by the Leave-One-Out (LOO) method. Let *d* be a dataset consisting of *N* observations: the number of folds in which to divide *d* will be equal, in the case of the LOO, to the number of observations n ($k = N$) so that at each iteration, only one sample will be used as a test set. This approach is recommended with small datasets.

In classification problems, the *stratified* variant for cross-validation is preferred as it preserves the proportions of each class, not only in the training set but also in the folds created to represent the test set.

Y-scrambling (also known as Y-randomization or response randomization) techniques should be used to rule out the possibility of chance correlation and to inspect for reliability and robustness by permutation testing: the dependent-variable vector, **y**-vector, is randomly shuffled, and a new model is trained using the initial independent-variable matrix **X**[11]. The process is repeated many (*N*) times, and the performance of the models trained with scrambled data is measured by calculating the average values of the validation metrics. The model is considered robust if the *N* models' statistical parameters, trained with scrambled data, show much lower performance[12].

Hyperparameter optimization[13] techniques, such as grid search, random search[14], and Bayesian optimization[15], represent optimization strategies for classification and regression models. By tuning the model's parameters, it is possible to find the parametric combination that best enhances the learning algorithm's performance. These techniques are particularly useful for algorithms that require the setting of two or more parameters simultaneously. The greater the number of parameters to be set, the greater the calculation times required to realize the technique.

The most common metrics for measuring the performance of classification tasks are:

- Accuracy (ACC). Equation ((1) is the proportion of true positives (TP) and true negatives (TN) among the total cases examined (P + N). That is the ratio of correct classifications to the total number of correct or incorrect classifications.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} \tag{1}$$

- **Precision**, **Positive Predictive Values** (PPV). Equation (2) describes the ability of the classifier not to label as positive samples that are negative. That is, given a positive prediction, how likely is the classifier to be correct. The best value is 1, and the worst value is 0.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

- **Recall**, **Sensitivity**, **True Positive Rate** (TPR). Equation (3) measures the proportion of positively predicted labels that are correctly identified as such. That is the ability of the classifier to find all the positives samples. The best value is 1, and the worst value is 0.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

- **True Negative Rate (TNR)**. Equation (4) measures the proportion of negative predicted labels that are correctly identified as such. That is, is the ability of the classifier to find all the negative samples. The best value is 1, and the worst is 0.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4}$$

- **Receiver Operating Characteristic** (ROC) **Curve**[16]. It shows a binary classifier model's ability to discriminate between positive and negative classes as its discrimination threshold varies from high to low. The true positive rate (TPR) is plotted against the false positive rate (FPR) at various thresholds to obtain the curve. An area under the curve (AUC) of the ROC curve of 1.0 represents a model that correctly made all predictions. An AUC of 0.5 represents a model that is as good as a random classification (**Figure 3**).

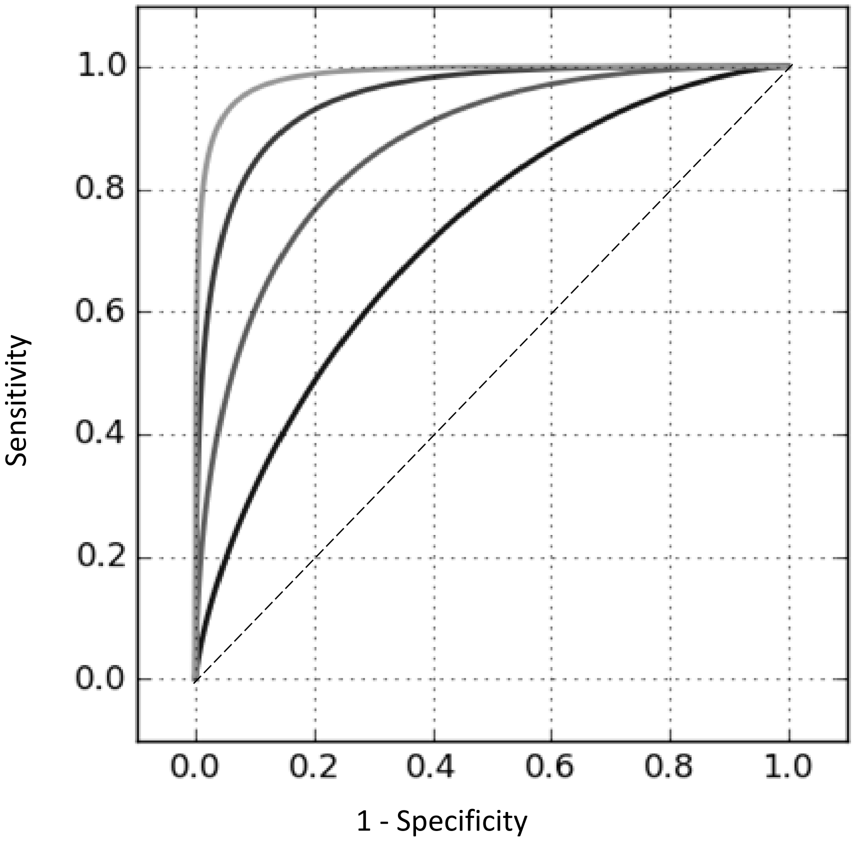


Figure 3. The ROC curve plotted with the TPR (sensitivity) against the FPR (1 – specificity). When AUC is 0.5 (dashed line), the model has no discrimination capacity to distinguish between positive class and negative class

- **Matthews Correlation Coefficient**[17] (MCC). Equation (5) is a measure of the quality of binary (two-class) classifications. The best value is 1, and the worst value is 0. It considers TP, FP, FN, TN, and is generally regarded as a balanced measure that can be used even if the classes are of different sizes. The MCC is a correlation coefficient value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. Only in the binary case does this relate to information about true and false positives and negatives.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{5}$$

The most widespread metrics for measuring the performance of regression tasks are:

- **Coefficient of determination** (*R*^{2} or *r*^{2}). Equation (6) represents the proportion of variance (of y) explained by the model's independent variables. It is a measure of how unseen samples are likely to be predicted correctly by the model through the proportion of explained variance. Therefore, it indicates the goodness of fit. The best possible score is 1, and it can be negative (because the model can be arbitrarily worse). A constant model will get an *R*² score of 0 if it always predicts the expected value of y, disregarding the input features. In equation (6), $\widehat{y_i}$ is the predicted value of the *i*-th sample and y_i is the corresponding true value for total n samples.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

- **Root Mean Square Error** (RMSE). Equation (7) is a measure of the difference between the predicted value and the true values. RMSE is the square root of the average of squared errors. It is a measure of accuracy, and it is scale-dependent. RMSE values have a range of $[0, +\infty]$ where a value equal to zero would indicate a perfect prediction. Therefore, a lower RMSE is better than a higher one. In equation (7), $\widehat{y_i}$ is the predicted value of the *i*-th sample and y_i is the corresponding true value for total n samples.

$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{\left(\frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \widehat{y_i})^2\right)} \tag{7}$$

1.3 Unsupervised Learning

Unsupervised learning algorithms process a training set that consists of a set of unlabeled input vectors **x** through evaluating the intrinsic and hidden structure of the data. Unlike supervised learning, only non-annotated examples are provided to the algorithm during learning, as the classes are not known a priori but must be learned automatically. The objective in such problems may be to discover groups of similar examples within the data, where it is called **clustering**, or to determine the distribution of data within the input space, known as **density estimation**, or to project the data from a high-dimensional space down to two or three dimensions for visualization known as **dimensionality reduction**.[4]

Unsupervised learning techniques work by comparing data and looking for similarities or dissimilarities. They are very efficient with numerical elements since they can use all the techniques derived from statistics, but they are less efficient with non-numerical data. These algorithms work correctly in data containing clear and identifiable sorting or grouping. If the data is endowed with an intrinsic ordering, the algorithms are still able to extract information. On the contrary, they can fail.

A typical example of these algorithms is found in search engines. Given one or more keywords, these programs can create a list of links referring to the pages that the search algorithm considers relevant to the search performed. The validity of these algorithms is linked to the usefulness of the information they can extract from the database. In the above example, it is linked to the relevance of the links with the searched topic.

1.3.1 Dimensionality reduction

Dimensionality reduction methods transform data from a high-dimensional space into a low-dimensional space representation that retains a subset of the original data's properties, ideally close to its intrinsic dimension (the number of variables needed for a minimal representation of the data). Working in high-dimensional spaces can be inconvenient for various reasons:

- Raw data are often sparse.
- Analyzing the data is usually computationally intractable.

Dimensionality reduction is standard in fields that deal with large numbers of observations and/or large numbers of variables, such as bioinformatics and cheminformatics.[18]

Dimensionality reduction algorithms can be classified into linear and non-linear.[18] Moreover, they can also be classified into feature selection and feature extraction methods.[19] Dimensionality reduction can be used for noise reduction, feature engineering, data visualization, cluster analysis, or intermediate steps to facilitate other analyses.

The primary technique for feature extraction is the Principle Component Analysis (PCA)[20]. PCA performs a linear mapping of the data to a lower-dimensional space so that the variance of the data in the low-dimensional representation is maximized. The lost information is usually regarded as noise.

1.3.2 Clustering

Clustering is the unsupervised partitioning of data into homogeneous groups (clusters). Items in each group are more like each other than items in another group. Therefore, the objective is the identification of homogeneous subgroups among a set of heterogeneous items. The definition of object and features in each clustering analysis depends on the hypothesis being tested. The dimensionality is defined by the number of features an object has rather than the number of objects clustered.

Clustering is a standard method for analyzing data sets in pharmaceutical sciences. The clustering objective is to gain insight into the underlying structure in the complex data, find basic patterns within the data, uncover relationships between molecules, biomolecules, biological entities, conditions, and use these discoveries to generate hypotheses and decide further biological experimentation. It is a fundamental analysis for understanding and visualizing the complex data acquired in high-throughput multidimensional biological experiments. The amount of data generated in pharmaceutical sciences is experiencing a massive scale-up. Extraction of relevant information is becoming increasingly challenging, and data analysis methods such as clustering are essential. Molecules in chemical clustering are usually encoded as physicochemical descriptors, fingerprints, or graph properties input vectors. The chemical space is estimated to be 10^{63} compounds in the context of computer-aided drug discovery. Therefore, clustering techniques can be useful to select promising subgroups inside a sizeable chemical library by discarding *a priori* the bulk of the dataset with either no pharmaceutical interest or characterized by redundancy in physicochemical and topological properties. Moreover, clustering analysis helps identify outliers, understand a particular functional group's behavior, and identify common scaffolds in each set of molecules.

There are many clustering algorithms and no single best method for all datasets. Typical cluster models include:

- Connectivity models - are based on distance connectivity (e.g., hierarchical clustering[21,22])
- Centroid models - represent each cluster as a single average vector (e.g., k-means[23,24])
- Distribution models - use statistical distributions for clusterization (e.g., gaussian mixture[25])
- Density models define clusters as connected dense regions in the data space (e.g., density-based spatial clustering (DBSCAN)[26,27])

1.4 Quantitative structure-activity (property) relationship

The most critical application of Machine Learning in pharmaceutical sciences is predicting chemical bioactivity and physical properties. This research field is known as quantitative structure-activity relationship (QSAR) modeling and quantitative structure-property relationship (QSPR) modeling, a well-established computational approach to chemical data analysis. It has found a broad range of physical organic and medicinal chemistry applications in the past 55+ years.[28]

QSAR/QSPR models are developed by establishing empirical, linear, or non-linear relationships between values of chemical descriptors computed from molecular structure and experimentally measured properties or bioactivities of those molecules, followed by applying these models to predict or design novel chemicals with desired properties.

The traditional areas of QSAR/QSPR modeling are drug discovery and development and chemical safety prediction.[29] Recent technological advancements in Machine Learning allowed the application of new algorithms, modeling methods, and validation practices to a wide range of pharmaceutical research areas outside of traditional QSAR/QSPR boundaries, including analytical chemistry, synthesis planning, nanotechnology, materials science, clinical informatics, biomaterials, and quantum mechanics.[28–30] Many publications have advanced the traditional QSAR modeling[28,29], such as prediction of biological activities and ADME/Tox properties, building on the successful use of QSAR modeling in chemical, agrochemical, pharmaceutical, and cosmetic industries[28]. However, new and exciting directions and application areas have also emerged, such as process chemistry[31,32], synthetic route prediction and optimization, and retrosynthetic analysis[28]. Thus, machine learning models have become an integral component of the drug discovery process, providing substantial guidance in planning experiments[29,33].

1.5 Carotenoid content in Goji berries

In the present work, machine learning methods were applied to extracts of *Lycium barbarum* L. fruits[34]. Fruits of *Lycium barbarum* L. have been of much interest due to its biological constituents' potential health benefits. The high level of carotenoids in these fruits offers protection against the development of cardiovascular diseases, diabetes, and related comorbidities. Two different selections of *Lycium barbarum* L., cultivated in Italy and coming from three discrete harvest stages, were subjected to two different grinding procedures and a simplified extraction method of the carotenoid component. CIELAB colorimetric analysis of the freshly prepared purees and HPLC-DAD analysis of carotenoid extracts were performed and compared. A significant carotenoid fraction, responsible for the characteristic orange-red color, makes Goji berries one of the richest carotenoid natural sources. Zeaxanthin dipalmitate, a molecule with a highly valuable biological role, was the most representative compound of this class. Machine learning techniques were applied for the analysis of the samples extracted. A QSPR model was developed through supervised learning methods. Linear correlations between carotenoid and zeaxanthin amounts with colorimetric features were defined and statistically validated. The encouraging results indicate that quick and economic colorimetric analysis, directly performed on homogenized samples and enhanced by machine learning, could enable the zeaxanthin and carotenoids quantity prediction. Thus, the final QSPR model provides a reliable tool to directly assess carotenoid content by performing cheap and routinely colorimetric analyses for the food industry.

1.6 Antimicrobial and antibiofilm activity of essential oils

Essential oil is a mixture of low molecular weight constituents that are responsible for its characteristic aroma. These constituents include terpenoid and non-terpenoid hydrocarbons and their oxygenated derivatives. It is well known that essential oil compositions can differ according to geographical region, seasonality, and extraction methodology[35–37]. Therefore, essential oil's chemical composition is heterogeneous with unique characteristics (chemotype) and specific effects on microorganisms.

A biofilm is a complex aggregation of a syntrophic consortium of microorganisms characterized by the secretion of an adhesive and protective matrix, composed of extracellular polymeric substances (a conglomeration of polysaccharides, proteins, lipids, and nucleic acids) that adheres to a biological (for example, roots of plants and epithelium of animals) or inert (for example, prostheses or rocks) surface[38,39]. The ability to form biofilms is a universal attribute of bacteria[39]. *Bacillus, Escherichia*, *Pseudomonas*, *Staphylococcus*, and *Streptococcus* are among the clinically relevant species belonging to genera forming biofilms[39].

Biofilm formation of bacteria is related to their colonization of new environments. A biofilm lifestyle is associated with a high tolerance to exogenous stress. Therefore, the treatment of biofilms with antibiotics or other biocides is usually ineffective at eradicating them. Biofilm formation is a major problem in many fields, from the food industry to medicine, and is the cause of persistent infections implicated in 80% or more of all microbial cases, releasing harmful toxins, and even obstructing indwelling catheters. The development of anti-biofilm agents is considered of significant interest and represents a key strategy as non-biocidal molecules are highly valuable to avoid escape mutants' rapid appearance. Considering these assumptions, the interest in developing new approaches for preventing bacterial adhesion and biofilm formation has increased. Antimicrobial and antifungal properties have been attributed to essential oils. The wide use of essential oils applies to aromatherapy, household cleaning products, personal beauty care, and natural medical treatments. Recently, several reports indicated *in vitro* efficacy of non-biocidal essential oils as a promising treatment to reduce bacterial biofilm production and prevent drug resistance[40–42].

Machine learning techniques were applied to the chemical composition of essential oils and correlated to the experimentally determined associated biofilm modulation influence that was either positive or negative. Quantitative Activity-Composition Relationships (QCAR) were developed through machine learning classification techniques to discover the chemical components mainly responsible for the anti-biofilm activity. QCAR models are an example of machine learning applied to investigate complex chemical mixtures.

Statistically, robust classification models were developed, and their analysis in terms of feature importance and partial dependence plots led to indicating those chemical components mainly responsible for biofilm production, inhibition, or stimulation for each studied strain, respectively. Model agnostic feature importance and partial dependence plots were used to find the marginal effect that each essential oil chemical component has on the predicted outcome of the binary classification models. From the results of these studies, it could be observed that each essential oil has a specific effect on biofilm formation, likely depending on its characteristics and unique chemical composition.[43,44]

In the last chapter, unsupervised learning analysis of clinical strains of bacteria that cause cystic fibrosis is reported[45]. The most severe infections reoccurring in cystic fibrosis, one of the most common lethal genetic disorders in the Caucasian population, are due to *S. aureus* and *P. aeruginosa*[46]. Intensive use of antimicrobial drugs to fight lung infections inevitably leads to the onset of antibiotic-resistant bacterial strains. New antimicrobial compounds should be identified to overcome antibiotic resistance in these patients. Therefore, an extensive study on 61 essential oils against a panel of 40 clinical strains of *S. aureus* and *P. aeruginosa* isolated from cystic fibrosis patients was conducted. To reduce the in vitro procedure and render the investigation as convergent as possible, unsupervised machine learning algorithms were applied to pick-up a fewer number of representative strains (clusters of strains) among the panel of 40. This approach allowed the rapid identification of three essential oils that strongly inhibit bacterial growth of all bacterial strains considered in this research. Interestingly, the antibacterial activity of essential oils was unrelated to each strain's antibiotic resistance profile.

2 Carotenoid content of Goji berries: CIELAB, HPLC-DAD analyses and quantitative correlation

2.1 Introduction

Fruits provide nutrients for humans but also prevent nutrition-related diseases. Degenerative illnesses, such as cardiovascular diseases, diabetes, and relative comorbidities, are among the leading causes of death in all industrialized countries. Thus, consumers are oriented toward consuming food with documented health properties in search of a better lifestyle to prevent these pathologies[47]. *Lycium barbarum* L. fruit (Goji berry, GB) represents the focus of many scientific studies aiming to evaluate its content in bioactive compounds that could improve health. Goji berries are traditional Asian food, and China is the largest producer in the world[48]. More than 70 different species of *Lycium* exist in nature, among which *Lycium chinense*[49] is more common than the most appreciated for its phytochemical composition[50,51] and its antioxidant and radical scavenging properties, *Lycium barbarum*.

GBs are used in traditional Chinese medicine for liver protection, antioxidant purposes. They are also recommended as food supplements for their promising antiaging and cancer preventive role, cardiovascular protection, and therapeutic activities on immune system functionality[52]. This fruit's health benefit potential demanded investigations of its chemical composition, thus leading to the identification of polysaccharides, monosaccharides, organic acids, proteins, flavonoids and derivatives, carotenoids, vitamins, and mineral salts[53,54]. Carbohydrates represent about 51% of berries components[55] among which the water-soluble polysaccharide fraction has received significant attention in the last years. Furthermore, arabinogalactan proteins have been identified as significant bioactive molecules for their hypoglycemic and hypolipidemic effect[52,56].

The antioxidant and protective role of GBs was evaluated by Oxygen Radical Absorbance Capacity (ORAC) and by radical scavenging activity and associated with the high content of phenylpropanoids and (iso)flavonoids (caffeic and chlorogenic acid, quercetin-3-O-rutinoside and kaempferol-3-O-rutinoside), coumarins, lignans[54,57–59].

Goji berries are one of the richest natural sources of carotenoids (CAR). Their significant carotenoid fraction is responsible for the characteristic orange-red color. One of the most common carotenoids found in goji berries is zeaxanthin (ZEA) in the form of zeaxanthin dipalmitate[48].

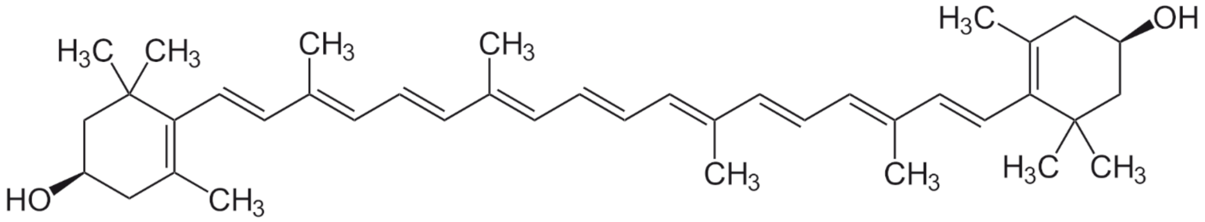


Figure 4. Zeaxanthin (ZEA).

ZEA accumulates in the macula densa of the retina and plays a protective role by preventing ultraviolet radiation degenerative effects. Lutein and zeaxanthin are protective agents towards age-related macular degeneration (AMD), and dried wolfberries are a rich source of such zeaxanthin esters[60]. AMD is a neurodegenerative disease that is considered the leading cause of acquired blindness. This progressive illness affects a significant number of elder people (over 55 years, about 9% worldwide) and has a multifactorial etiology: genetic, environmental risk factors, smoking, and dietary habits play a significant role[61,62]. A regular daily intake of fruits and vegetables that are rich sources of carotenoid antioxidant pigments and many other bioactive molecules has been associated with a reduced risk of chronic and degenerative diseases[63]. Several factors could influence the bioavailability of these components (chemical nature, food matrix, human metabolism, absorption efficiency in the intestinal lumen), and consequently, their efficacy in health promotion[63]. The fact that hydrophilic polysaccharides and lipophilic carotenoids exert different functions such as generic protection towards oxidation, type-2 diabetes, inflammation, cancer, and that could help prevent specific illnesses, increase the interest for GBs as a food supplement. Type-2 diabetes and cardiovascular diseases could also have some critical roles in AMD progression and diseases associated with retinopathies. Hypoglycemic polysaccharides prevent the onset and the progression of diabetes[56]. Their combined action with ZEA, active in the macular protection, could provide a synergic effect in the blindness prevention. The present work's objective was to monitor the carotenoid content in Goji berries cultivated in Italy, evaluating the differences among varieties, harvesting periods, seasons, and extraction procedures. CIELAB colorimetric and quali-quantitative HPLC-DAD analyses were performed and the obtained data were statistically analyzed to build correlation models aimed to predict CAR and ZEA contents directly from colorimetric measurements.

2.2 Material and Methods

2.2.1 Materials

Ethanol (≥ 96%), double-distilled water, cyclohexane RPE, methanol RS, and acetone RS for HPLC were purchased from Carlo Erba (Milan, Italy). HPLC-grade glacial acetic acid and ethyl acetate were obtained from Fluka (Milan, Italy). GBs were generously gifted by Azienda Natural Goji® and were harvested at different commercial harvesting periods 1-5 (2015: July 6th 1; July 23rd, 2; August 3rd, 3; and 2016: July 26th, 4; August 4th, 5) in Fondi (Latina province, Lazio region, Italy) based on their stage maturity as determined by the producer. They were "Poland" and "Wild" varieties, P and W, respectively. Ten different samples (P1-P5 and W1-W5) were then collected, quickly frozen at -80 °C and stored at -18 °C, until the analyses were performed. Zeaxanthin dipalmitate standard (purity ≥ 98%) was purchased from Extrasynthese (Lyon, France).

2.2.2 Samples preparation

The defrosted GBs were washed and wiped up on paper towels at room temperature. Afterward, they were ground and homogenized with two different procedures: at room temperature for 2 minutes using a domestic mixer at 16,000 rpm (D samples) or by a T18 Ultraturrax® homogenizer (IKA®, Staufen, Germany) at 10,000 rpm (U samples). The procedure steps were conducted cautiously to reduce the loss of pigments due to GBs lability. The resulting fruit purees were further divided into two aliquots: one for the CIELAB colorimetric analysis and the second for the extraction procedure leading to a panel of 20 experiments (P1D-P5D, P1U-P5U, W1D-W5D, and W1U-W5U - **Table 1** and **Table 2**).

Table 1. P1D-P5D, **P1U-P5U**, **W1D-W5D,** and **W1U-W5U** Residue isolated from the cyclohexane fraction (ORG), zeaxanthin (ZEA), carotenoids (CAR), and yields from Goji berry (GB) extractions. The ratio between ZEA and CAR is also reported. All data are expressed as mg per gram (g) of GBs dry weight. Sample names were compiled merging the selection **P** or **W**, the number of harvesting (1-5), and the homogenization technique (**D** or **U**) as reported in the text. Mean values were from four different experiments (errors in the range of 5-10% of the reported values).

SAMPLE	ORG	ZEA	CAR	ZEA/CAR ratio (%)
P1D	20.85	5.03	6.40	78.6
P2D	18.56	2.63	3.92	67.3
P3D	21.13	1.90	2.61	73.1
P4D	20.17	4.49	6.00	74.9
P5D	14.50	2.74	1.82	72.8
P1U	29.66	4.02	5.04	79.7
P2U	36.83	2.28	5.94	77.3
P3U	33.91	2.45	3.32	73.8
P4U	30.53	4.30	5.64	76.3
P5U	21.89	2.94	3.95	74.3
W1D	20.27	5.54	6.86	80.7
W2D	20.36	3.25	5.66	57.7
W3D	18.64	2.16	3.93	55.1
W4D	20.18	4.99	6.38	78.3
W5D	13.81	2.54	1.83	69.2
W1U	30.51	4.26	5.87	72.8
W2U	43.61	2.56	9.18	56.2
W3U	30.70	1.92	6.41	30.1
W4U	25.36	3.94	4.94	79.7
W5U	20.00	2.27	3.26	69.8

Table 2. Colorimetric data of the Goji berry **P1D-P5D**, **P1U-P5U**, **W1D-W5D,** and **W1U-W5U** samples. Mean values were from four different experiments (errors in the range of 1-2% of the reported values).

SAMPLE	L*	a*	b*	C*	hab
P1D	37.63	21.78	18.92	28.85	40.98
P2D	40.47	24.66	23.89	34.33	44.10
P3D	40.27	24.54	23.43	33.93	43.66
P4D	39.61	26.17	20.88	33.48	38.58
P5D	37.93	22.50	18.50	29.13	39.42
P1U	40.54	26.26	23.27	35.01	41.65
P2U	39.18	22.58	21.43	31.13	43.51
P3U	42.18	26.90	26.86	38.01	44.95
P4U	39.03	25.90	20.82	33.23	38.79
P5U	38.86	24.30	19.85	31.38	39.25
W1D	38.22	22.06	19.34	29.34	41.25
W2D	39.71	24.04	22.00	32.59	42.46
W3D	40.42	23.58	23.31	33.16	44.67
W4D	38.34	23.89	19.38	30.76	39.04
W5D	35.65	16.81	14.71	22.34	41.17
W1U	44.44	30.54	28.95	42.08	43.47
W2U	43.86	29.20	29.16	41.27	44.96
W3U	43.45	27.27	28.12	39.17	45.87
W4U	39.46	26.49	20.79	33.65	38.16
W5U	39.09	23.67	19.77	30.85	39.87

2.2.3 Extraction of organic fractions

About 5 g of the obtained purees were extracted for 3 hours with 15 mL of a hydroalcoholic mixture (ethanol:water 70:30 v/v; water was previously acidified with 0.5% acetic acid) and 15 mL of cyclohexane, at room temperature and in the dark, under stirring. An upper organic phase, an intermediate hydroalcoholic, and a lower solid phase were present. The resulting suspension was centrifuged at 12000 g for 10 minutes at 4 °C, and the supernatant was collected and dried by a rotary evaporator at reduced pressure and 40 °C. Storage of residues (ORG) was reduced to a minimum, and samples were protected from light, heat, and air exposure. Finally, an aliquot of 5 mg was weighed and dissolved in ethyl acetate (5 mL) for the subsequent HPLC-DAD analyses.

2.2.4 HPLC-DAD analysis

Each dried organic fraction was subjected to HPLC analysis employing a Perkin-Elmer apparatus equipped with a series LC 200 pump, a series 200 diode array detector, and a series 200 autosampler. Data acquisition and processing were carried out with a Perkin-Elmer Totalchrom software. The chromatographic separation was performed using a Luna (Phenomenex) RP18 column (250×4.6 mm, i.d. 5 µm). The mobile phase (flow rate of 1 mL/min) consisting of acetone (solvent A) and methanol (solvent B) in 35 minutes was changed from 55% A and 45% B to 80% A and 20% B. 20 µL were injected setting at 450 nm the detector wavelength[64]. Peak assignments were made based on their ultraviolet-visible spectra, co-chromatography respect to commercial standards, when available, and by comparison with elution order as reported in other published studies[65]. ZEA was quantified by an external-matrix matched calibration method on the basis of the area ratios respect to the pure chemical standard. CAR was calculated as the sum of all the identified chromatographic peaks. The concentrations were reported as mg/100 g of dry fruit (**Table 1**).

2.2.5 Colorimetric analysis

CIELAB parameters (L*, a*, b*) were determined directly on the homogenized samples using a colorimeter X-Rite SP-62 (X-Rite Europe GmbH, Regensdorf, Switzerland), equipped with D65 illuminant and an observer angle of 10° (**Table 2**). Cylindrical coordinates C^*_{ab} and h_{ab} are calculated from a* and b* by equations (8) and (9)[66].

$$C^*_{ab} = (a^{*2} + b^{*2})^{\frac{1}{2}} \tag{8}$$

$$h_{ab} = \tan^{-1}\frac{b^*}{a^*} \tag{9}$$

2.2.6 Statistical analysis

Data analysis, calculations, and simulations were performed employing Python programming language (version 3.6.4, Python Software Foundation, https://www.python.org/). All calculations were done on a blade server (dual socket Intel Xeon E5520 2.27GHz CPUs and 24 GB DDR3 RAM) with a Debian GNU/Linux 9.3 operating system. Several Python libraries were used: the interactive shell IPython[67] (version 6.2.1), NumPy[68] (version 1.13.3), Matplotlib[69] (version 2.1.1) Pandas[70,71] (version 0.22.0) for software prototyping and development, interactive data analysis and visualization, optimizations, and simulated annealing simulations. The Huber Loss Regression (HLR) model and its validation were performed employing the Machine-Learning library Scikit-learn[72] (version 0.19.1). A cross-validation procedure assessed models' robustness (internal predictivity) with the leave one out (LOO) method. Models' chance correlations were evaluated via the Y-Scrambling approach, using 1000 iterations.

2.3 Results and discussion

2.3.1 Water content detection

About 5 g of freshly defrosted GBs samples were dried to constant weight, and their water content was assessed to be about 78%, approximately 8% less than those previously reported[60,65]. This determination allowed us to compare our experimental results, obtained by fresh samples, with literature data often referred to as dry weight[73].

2.3.2 Homogenization and extraction

During homogenization, fresh tissues are destroyed, leading to the release of enzymatic and acid components that might alter both the oxidative state and the trans-cis isomerization of the carotenoid fraction[74]. Therefore, two different techniques were applied to simulate the domestic and industrial procedures before optimizing the carotenoid extraction. Despite their lipophilic nature, carotenoids are within the aqueous biological matrix, so that their quantitative extraction requires hydrophilic solvents that could penetrate inside the tissues and lipophilic solvents to dissolve them. This issue is often addressed by a preliminary extraction followed by a liquid/liquid partitioning[74,75]. In this report, to the best of our knowledge, for the first time, an extraction method based on a double phase system was used so that polar and carotenoid fractions were simultaneously extracted and purified from each other. The resulting mixtures were then submitted to a centrifugation step to split the two phases better. In **Table 1** are reported the extraction yields of the residue isolated from the cyclohexane fraction (ORG) in the twenty analyzed samples. All extraction experiments were performed in quadruplicate (all the errors fall in the range of 5-10%). Average values range between 0.3 and 1.0% in fresh weight and 1.4 and 4.4% in dry weight of Goji fruits. The highest value was found in W2U, and it is about three-fold respect to the lowest value found in P5D (**Table 1**). ORG was found more abundant in extractions performed by the U procedure (29.9 mg/g dry weight), more than 1.5-fold higher respect to the D method (about 19.2 mg/g dry weight). Considering the two GBs varieties (**Figure 5**), no differences were found in the mean values (P = 24.6 mg/g dry weight, W = 23.2 mg/g dry weight). Differences were shown between seasons 2015 (exp. 1-3, 26.3 mg/g dry weight) and 2016 (exp. 4, 5; 20.3 mg/g dry weight) with a 30% more in 2015. So, the more marked variations are shown between the Series D and U, suggesting that Ultraturrax® homogenization enables a better solvent permeation and limits the lipophilic component degradation respect to the domestic mixer application.

2.3.3 HPLC analysis

The residues from the organic phases were further characterized by HPLC-DAD analysis at 450 nm to determine the carotenoid components (CAR). Among them, the presence of ZEA, as the most representative element of the CAR fraction of Goji berries, was confirmed by a pure reference standard and literature data[76]. The samples W1D and W3D (maximum and minimum value of zeaxanthine dipalmitate, respectively) were reported as example chromatograms. From eight to thirteen peaks were detected, only some of which were tentatively identified by comparison with previously published data as zeaxanthin, β-cryptoxanthin, and antheraxanthin mono and diacylates[65,77]. The results obtained by the performed analyses are shown in **Table 1** and in **Figure 5**, where the mean values obtained by the different series were compared. CAR content was calculated as the sum of all the peak areas revealed at 450 nm and expressed as ZEA equivalents. Taking into account this approximation, ZEA ranged between 55% and 81% of the total carotenoids, with the only exception of sample W3U (**Table 1**). These data are consistent with Karioti et al. (2014), which reported zeaxanthin dipalmitate as principal component (82 and 87%) of carotenoids extracted by commercial samples of dried GBs[64]. No differences were shown between varieties P and W. Comparing U *vs* D and 2015 *vs* 2016, the lowering of ORG corresponds to a lower content of CAR, but a higher content of ZEA was detected as if ZEA showed higher stability towards other carotenoid components. The ZEA (3.11-3.49) and the CAR contents (4.27-5.23 mg/g dry weight) of all the series were comparable. Therefore, according to the obtained data, the domestic homogenization (D method), although supplying a lower ORG yield, accounts for a higher ZEA content. The best yields were achieved in samples deriving from harvesting 1 (2015) and 4 (2016), corresponding to the commercial stage maturity of the collected fruits, also if the harvesting dates do not correspond strictly (July 6th vs July 26th). The highest ZEA content, obtained by D procedure, ranged between 4.5 and 5.5 mg/g of dry matter, and over-lapped in the two different selected varieties P and W and in the two considered harvesting dates (1: 2015 and 4: 2016).

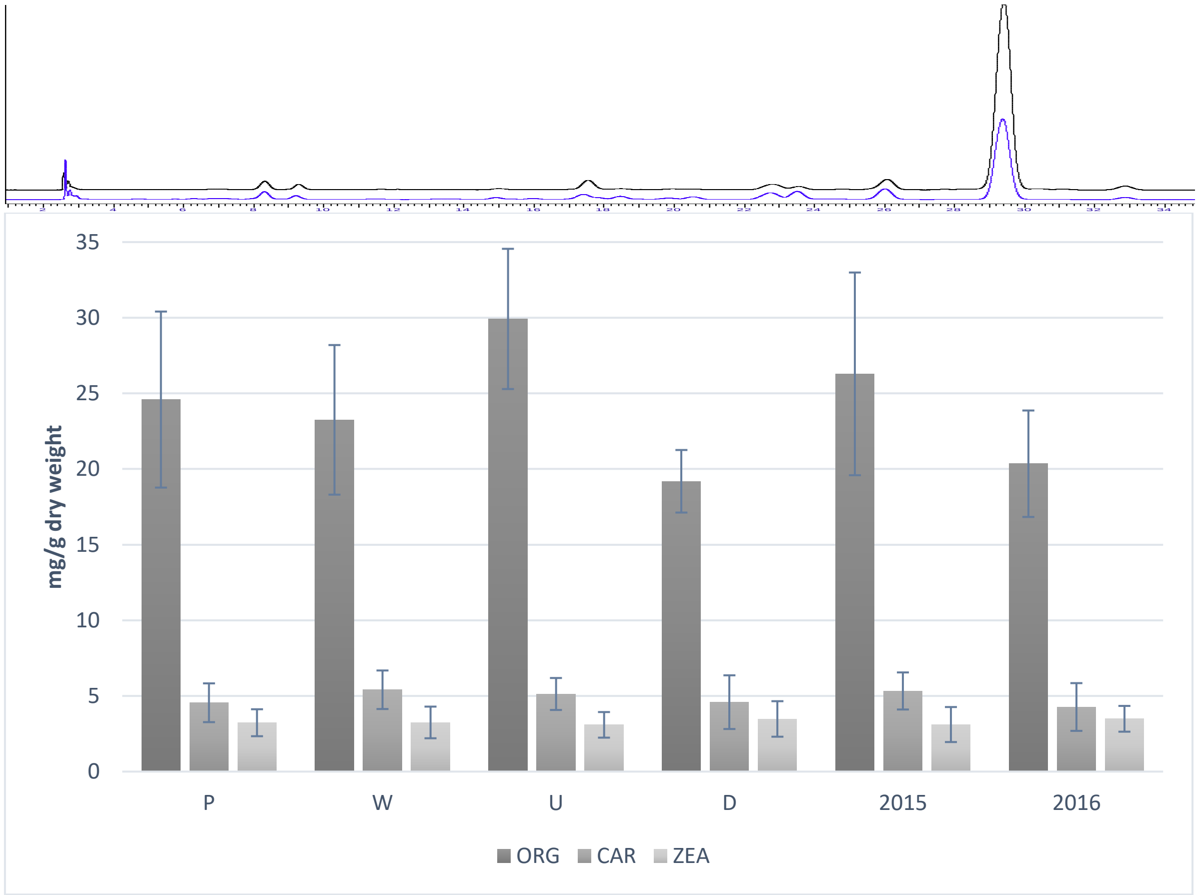


Figure 5. Example chromatograms (W1D and W3D) of carotenoid content and comparison among mean contents found for the Series P (Polonia), W (Wild), D (domestic), U (Ultraturrax®), harvested in 2015, and harvested in 2016.

2.3.4 Color analysis

Color plays a fundamental role in consumer choices. It is associated with quality and genuineness and could be correlated with the presence of a characteristic chemical profile (pigments). The possibility to analyze food matrices as turbid juices, jellies, or homogenates, without any destructive treatment, represents a further advantage aimed to clarify what could alter the pigment content. The CIELAB data, obtained by the colorimetric analyses of the twenty Goji samples (**P1D**-**P5D**, **P1U**-**P5U**, **W1D**-**W5D,** and **W1U**-**W5U**) were reported in **Table 2**, and the resulting reflectance profile curves were shown in **Table 2**. The luminance *L** ranges between a minimum of 35.65 and a maximum of 44.44. The redness *a** values range between 16.81 and 30.54. The yellowness *b** ranges between 14.71 and 29.16 with relevant differences. Results account for a comparable contribution to the color of the two parameters *a** and *b**. The "Wild" variety showed the lowest and the highest values (**W5D**: *L** 35.65, *a** 16.81, *b** 14.71; **W1U**: *L** 44.44, *a** 30.54, *b** 28.95). Samples **W1U**, **W2U,** and **W3U**, all coming from the 2015 season, reached the highest intensity of orange color. Therefore, data did not show differences between the two varieties, but rather among the different harvesting dates and between the two different homogenization techniques, thus confirming the workup's preeminent role in preserving pigments and/or allowing a complete extraction from tissues. The above reported HPLC data also confirm this. Higher data for *L** (53), a* (30), and b* (40) were assessed directly on fresh Goji berries, without homogenization treatment[78]. The same values lowered down to 32, 12, and 5, respectively, after drying at 70 °C. In some reports[79–81], a correlation was found between HPLC or other quantitative analysis and color analysis of different food matrices. Several factors, such as fruit ripening, climatic conditions, varieties, and workup, could profoundly influence the pigment composition and, consequently, the color parameters, but only a few data have been published investigating carotenoids containing matrices such as wheat[79], orange juices[80], and corn[82]. In a study conducted by Humphries et al.[79], the results show a positive correlation between *b** and lutein content in the analyzed wheat samples. The comparison between HPLC and colorimetric analysis made by Kljak et al.[82] on Zea mays showed that *b** is directly correlated with lutein, β-cryptoxanthin, and β-carotene, whereas the pale orange colored zeaxanthin greatly influences *a**. Globally, the carotenoids increase determined a color intensification that provided a correlated increment of *C** and L* decrement. The high values of *a** found in our samples suggested that a significant contribution was due to the color nuance by the zeaxanthin dipalmitate structure. Together with the other carotenoids, its content could be responsible for the total color appearance, as depicted by the data analysis.

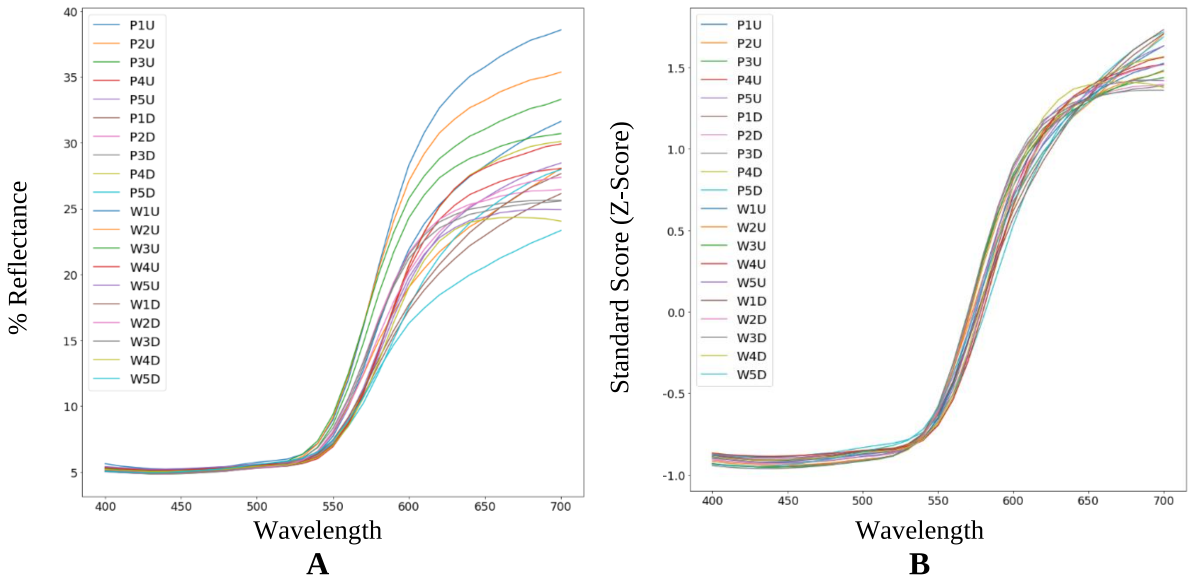


Figure 6. Reflectance curves of the Goji berries homogenized samples. Wavelength vs Reflectance percentage (Panel A) and Wavelength vs Standardized Reflectance percentage (Panel B).

2.3.5 Data correlation analysis

Three subsets of data types were available for the twenty data points (**P1D**-**P5D**, **P1U**-**P5U**, **W1D**-**W5D,** and **W1U**-**W5U**) herein listed: the first was derived from colorimetric experiments (L^*, a^*, b^* C^*, h_{ab}), the second from quantitative analyses (ZEA, CAR and ORG amounts), and the third from spectroscopic experiments. Pearson correlation coefficients (r) were calculated among ZEA or CAR and every colorimetric feature to seek for correlations among either ZEA or CAR content and colorimetric parameters. The only acceptable results, aside from the expected correlation among CAR and ZEA amounts (0.76), was an indirect correlation among h_{ab} and ZEA (-0.61). Reflectance values were recorded between 400 and 700 nm. Acceptable *r* coefficients, ranging between 0.5 and 0.6, were obtained for data collected at 400-480 nm with the maximum at 430 nm. For all the analyzed samples in this range, the reflectance percentages seemed quite similar when plotted independently (**Figure 6*A***). In contrast, a new profile was obtained by scaling the reflectance values (**Figure 6B**), highlighting some reflectance differences that could account for the found direct correlations.

2.3.6 Predictive linear model

A common approach to the analysis of a sample of data is to seek linear correlations between the variables that describe the sample and the dependent property of the sample itself (i.e., the concentration). Predictive linear models were built employing the Scikit-learn library[72]. Colorimetric features were normalized to zero mean and unit variance by subtracting the mean from each feature and by dividing the values of each feature by its standard deviation:

$$x' = \frac{(x - \bar{x})}{\sigma} \tag{10}$$

where x' is the standardized feature vector, x is the original feature vector, $\bar{x}$ is the mean, and σ the standard deviation. Due to the lack of satisfactory univariate linear regression models, a multivariate linear regression approach was adopted. A heuristic strategy to select the best features combination was set up. Simulated annealing[83] features elimination (SAFE) algorithm in conjunction with Robust Linear Regression[84] and Leave-One-Out (LOO) cross-validation was implemented in the Python programming language. The SAFE approach led to the optimized final models with 10 and 9 selected features for ZEA and CAR contents as dependent variables, respectively (**Figure 7**).

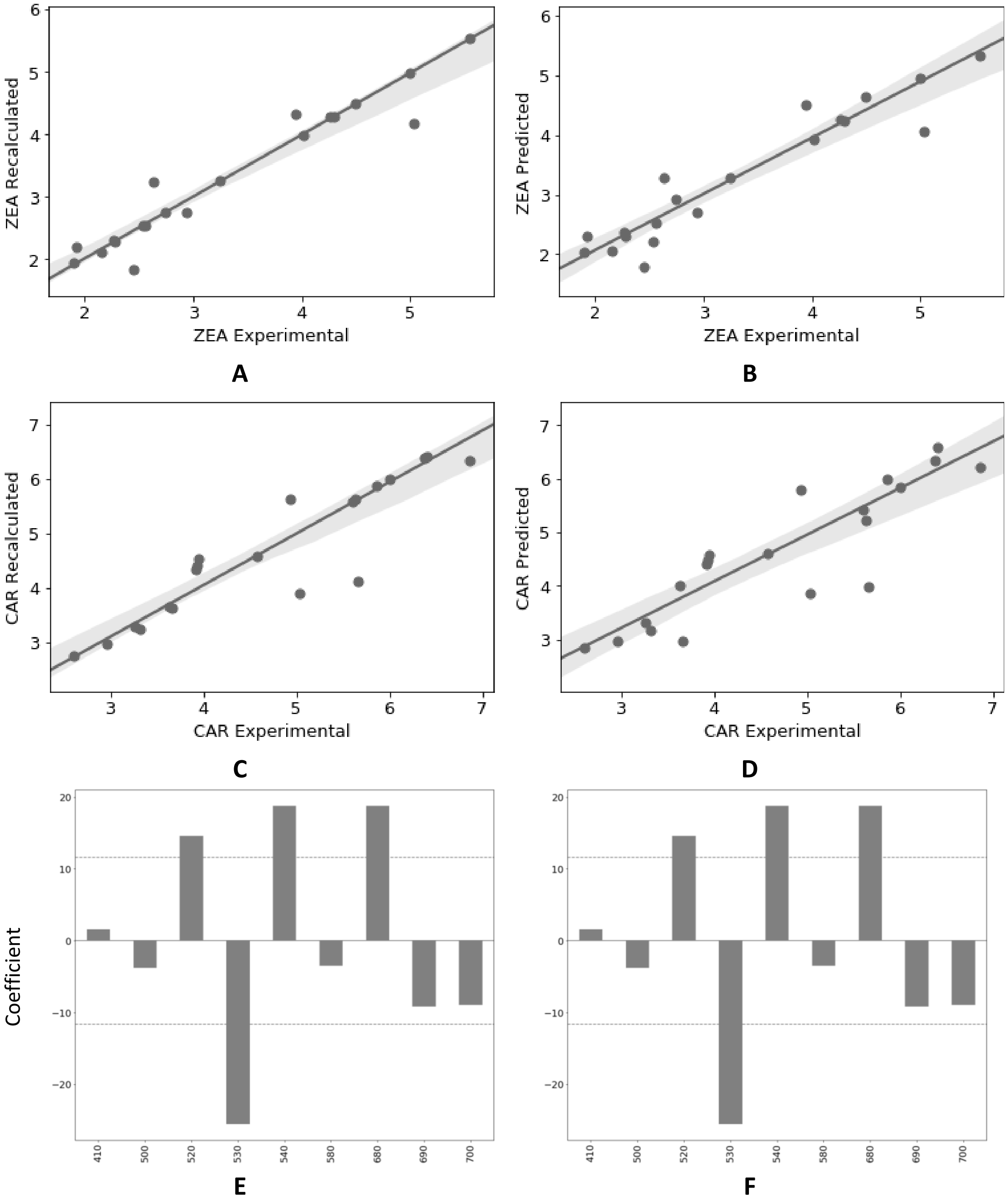


Figure 7. Experimental vs Recalculated/Predicted Concentrations of ZEA (Panels A, B) and CAR (Panels C, D). Plots were generated employing Seaborn Python library and Regression Coefficients of the ZAR model (Panel E) and CAR model (Panel F). Horizontal dashed lines correspond to the average of the coefficient's absolute values.

The robust regression predictive models were built using the Huber loss function (Huber Regressor), as implemented in scikit-learn. Models robustness and chance correlation absence were assessed through Leave One Out cross-validation and Y-scrambling procedures.

The resulting final models were endowed with good statistical coefficients (r^2, RMSE, q^2, RMSE$_{LOO}$), and the Y-scrambling procedure showed the absence of chance correlation (**Table 3**). Equations (11) and (12) define a sort of calibration curves that enable the direct quantity pre-diction of ZEA or CAR, respectively. Among the two models, the one for ZEA is slightly more precise (r^2 = 0.931) and more accurate (RMSE = 0.359) than CAR one. Both models are robust (see q^2 values in **Table 3**) and reliable as in the Y-scrambling validation, the mean $r^2_{Y\text{-}S}$ and $q^2_{Y\text{-}S}$ values are lower than the corresponding unscrambled models.

$$y_{zea} = \beta_0 + \beta_1 * x'_h + \beta_2 * x'_{400} + \beta_3 * x'_{410} + \beta_4 * x'_{490} + \beta_5 * x'_{520} + \beta_6 * x'_{530} + \beta_7 * x'_{600} + \beta_8 * x'_{610} + \beta_9 * x'_{660} + \beta_{10} * x'_{690} \tag{11}$$

$$y_{car} = \beta_0 + \beta_1 * x'_{410} + \beta_2 * x'_{500} + \beta_3 * x'_{520} + \beta_4 * x'_{530} + \beta_5 * x'_{540} + \beta_6 * x'_{580} + \beta_7 * x'_{680} + \beta_8 * x'_{690} + \beta_9 * x'_{700} \tag{12}$$

The regression models' coefficients of (11) and (12) directly measure the features' relative importance, as they were fitted using standardized parameters. Therefore, independent variables (features) with larger absolute values significantly affect the dependent variables (ZEA and CAR). The average absolute values (AAV) of coefficients in **Table 3** were calculated to be 8.17 and 11.63 for ZEA and CAR models, respectively. Standardized reflectance wavelength percentages at 600, 660, and 690 nm for ZEA model (equation (11) and 520, 530, 540, and 680 nm for CAR model (equation (12) displayed absolute coefficients higher than the respective AAV suggesting that ZEA is more sensible to reflectance measured in the narrow wavelength range (600-690 nm) differently from CAR which is described by a broader range (520-680 nm). These data indicate that it is possible to estimate ZEA content in CAR directly using equations (11) and (12).

Table 3. ZEA and CAR regression settings (estimator and optimized parameters) and statistical coefficients ***r^2***[a], **RMSE**[b], ***q^2***[c,] and **RMSE$_{\mathrm{LOO}}$**[d]. Below are also listed the selected features, and their regression coefficients and intercepts for the simulated anneal optimized ZEA and CAR models.

	ZEA	**CAR**
Algorithm	HR[e]	HR
epsilon[f]	1.5	1.4
alpha[g]	0.001	0.001
r^2	0.931	0.836
RMSE	0.295	0.510
q^2_{LOO}	0.897	0.770
RMSE$_{\mathrm{LOO}}$	0.359	0.603
$r^2_{\mathrm{Y\text{-}S}}$[h]	0.453	0.371
$q^2_{\mathrm{Y\text{-}S}}$[i]	-1.814	-1.616

Selected features and regression coefficients

#	**Name**	***$\overline{x}$***[k]	***σ***[l]	**Coefficient**	**Name**	***$\overline{x}$***[k]	***σ***[l]	**Coefficient**
β_0	intercept			3.29	intercept			4.66
β_1	h	41.90	2.50	5.38	410	5.14	0.14	1.57
β_2	400	5.23	0.15	-4.05	500	5.46	0.12	-3.80
β_3	410	5.14	0.14	7.47	520	5.65	0.14	14.58
β_4	490	5.35	0.12	-4.16	530	5.95	0.19	-25.53

β_5	520	5.65	0.14	6.53	540	6.55	0.38	18.79
β_6	530	5.95	0.19	-7.12	580	15.57	2.53	-3.47
β_7	600	21.00	3.22	-10.14	680	28.08	3.72	18.76
β_8	610	22.82	3.38	2.35	690	28.38	3.76	-9.18
β_9	660	27.21	3.63	20.55	700	28.66	3.82	-8.97
β_{10}	690	28.38	3.76	-13.93				

[a]: squared correlation or determination coefficient

[b]: root mean squared error of estimation

[c]: cross-validated squared correlation coefficient

[d]: cross-validated root mean squared error of prediction

[e]: Huber Regressor

[f]: epsilon is the parameter that controls the number of samples classified as an outlier

[g]: alpha is the regularization parameter

[h]: r^2 mean value from Y-scrambling model validation

[i]: q^2 mean value from Y-scrambling model validation

[k]: $\bar{x}$ mean of the feature vector

[l]: σ standard deviation of the feature vector

2.4 Conclusions

Two different Goji berry selections, cultivated in Italy, were submitted to an analytical evaluation that accounts for other homogenization techniques and different maturation stages. A new, rapid, cheap, and simple workup was developed and presented for the first time, as a one-step procedure for the extraction and purification of carotenoid bioactive components. The extraction method herein reported allowed to obtain a high amount of ZEA (about 5.5 mg/g dry weight of the analyzed Goji berries). Its pharmacological activity in the prevention of the ADM could endorse GBs as a suitable matrix in functional food or food supplement preparations. The quantitative HPLC-DAD analysis of the overall carotenoid content and the most representative xanthophyll zeaxanthin dipalmitate agreed with the colorimetric CIELAB analysis performed on the homogenized samples. Results allowed to highlight differences between the two cultivars and the harvesting periods and, even more noticeable, between the two applied homogenization techniques. Data obtained from the compared simulating domestic mixer and industrial Ultraturrax® high-speed homogenization underlines that this passage plays a predominant role in the preservation and subsequent extraction of the colored and bioactive components, so that the colorimetric analysis can show different characteristics. Thus, it is relevant both for the consumer choices, primarily based on food color and for industrial evaluations on the nutritional value of eventual food-derived products. Linear correlations between CAR or ZEA amounts with colorimetric parameters were defined and statistically validated. The encouraging results indicate that quick and economic colorimetric analysis, directly performed on homogenized samples, could enable the ZEA and CAR quantity prediction for commercial purposes. Further experiments are in due course to experimentally validate the models.

3 Antimicrobial and antibiofilm activity and machine learning classification analysis of essential oils from different Mediterranean Plants against *Pseudomonas aeruginosa*

3.1 Introduction

The extraordinary ability of bacteria to colonize new environments is undoubtedly related to biofilm formation. Biofilm lifestyle is associated with a high tolerance to exogenous stress: consequently, the treatment of biofilms with antibiotics or other biocides is usually ineffective at eradicating them. Biofilm formation is inevitably a significant problem in many fields, ranging from the food industry to medicine. In medical settings, biofilms are the cause of persistent infections implicated in 80% or more of all microbial cases, releasing harmful toxins and even obstructing indwelling catheters[85]. Bacteria of clinical relevance, such as *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Acinetobacter baumannii*, among others, proliferate on medical devices and form biofilms which provide them with up to 1000 times more effective resistance and tolerance to antibiotics in comparison with their planktonic forms[86].

P. aeruginosa is a common Gram-negative bacillus, able to adapt and survive in unfavorable environmental conditions, including minimal nutritional sources. It can cause disease in plants and animals, as well as humans. *P. aeruginosa* is a multidrug-resistant pathogen recognized for ubiquity, intrinsically advanced antibiotic resistance mechanisms, and association with serious illnesses, especially hospital-acquired infections, such as ventilator-associated pneumonia (VAP)[85] and various sepsis syndromes[87]. Severe infections caused by *P. aeruginosa* often occur during existing diseases or conditions, most notably cystic fibrosis and traumatic burns. In spite of the progress of antimicrobial therapies, infections by *P. aeruginosa* can still cause a mortality percentage range between 18% and 61% of cases[88,89]. The significant impact of *P. aeruginosa* infection is mainly due to its capability to form biofilm[90].

Once firmly established, the biofilm can be very difficult to eradicate as the bacteria are embedded in a self-produced polymeric substance, providing low susceptibility to conventional antimicrobial agents[91] and host defense cells of the immunologic system and resulting in chronic infections[92].

Considering these assumptions, the interest in the development of new approaches for the prevention of bacterial adhesion and biofilm formation has increased. Therefore, the development of anti-biofilm strategies is of significant interest and currently constitutes an important field of investigation in which non-biocidal molecules are highly valuable to avoid the rapid appearance of escape mutants[93]. Thus, the rationale of this study was to search for new antimicrobials that have the power to inhibit virulence instead of bacterial growth; such a choice may impose a weaker selective pressure for the development of antibiotic resistance to current antibiotics.

Compounds of natural origin still provide a high number of interesting structures, even in this era of combinatorial chemistry. Essential oils (EOs) represent a group of antimicrobial agents, which are complex mixtures of volatile secondary metabolites[94,95]. Essential oils (Eos) show antimicrobial and antifungal properties and are also largely used in various cultures for therapeutic and health purposes. The wide use of EOs applies to aromatherapy, household cleaning products, personal beauty care, and natural medical treatments. Furthermore, EOs may synergically enhance the antimicrobial potencies of some drugs[96,97]. Several EOs and phytochemicals have been reported to inhibit biofilm formation by bacteria and fungi[40,41], and their effects on *P. aeruginosa* have been studied[98,99]. Taking into account the same plant variety, EO composition can differ according to geographical region, seasonality, and extraction methodology[35–37].

This study reports chemical composition, antibacterial and anti-biofilm activity against *P. aeruginosa* of 89 different EOs obtained from 3 other plants harvested in different seasons and conditions: *Calamintha nepeta* (L.) Savi subsp. glandulosa (Req.) Ball (CG)[35], *Foeniculum vulgare* Miller (FV)[37], and *Ridolfia segetum* Moris (RS). Furthermore, quantitative activity-composition relationships (QCAR) were developed through machine learning classification approaches to discover the chemical components mainly responsible for the anti-biofilm activity.

3.2 Material and Methods

3.2.1 Plants

Fresh aerial parts of CG[35], FV[37], and RS were collected in a wild area about 15 km from Tarquinia city (Province of Viterbo, Italy), in the archaeological zone near the Etruscan temple Ara della Regina (42°15′31.8″ N, 11°48′08.7″ E). The material was collected in summer and early autumn periods of the year 2015 and monitored for four (CG) and three (FV) months, from July to October, thus covering pre-, during-, and post-flowering periods. Regarding RS, this is an annual summer plant, completing its life cycle within June/July. Hence, this species has been monitored in different extraction times, not periods of harvesting. CG oils were obtained directly from fresh plant material, while FV and RS were air-dried in a shady place for 20 days. Voucher specimens have been deposited in the Department of Drug Chemistry and Technology at Sapienza University of Rome, Italy. Taxonomic identification of the chosen species was conducted according to the official European flora and the National Italian one

3.2.2 Essential Oils Extraction

Essential oils have been isolated, as previously reported, by direct steam distillation using a 62 L steel distillation apparatus (Albrigi Luigi E0131, Verona, Italy)[35,36,100]. Briefly, plant materials (about 1.5 kg) were subjected to fractioned steam distillation[36], collecting EOs at six interval times of 1, 2, 3, 6, 12, and 24 hours. In the case of RS a seventh fraction was collected after 30 hours.

At each fraction, the oil/water double phase was extracted three times, with 20 mL of diethyl ether. The organic layers were dried over anhydrous sodium sulfate (Na_2SO_4), filtered, and deprived of the solvent in vacuo to furnish the final EOs, which were stored in the freezer in tightly closed dark vials until further analysis.

Besides, to simulate parallel continuous EO extraction for 2, 3, 6, 12, 24, and 30 hours, mixtures were prepared by adding different amounts of diethyl ether to each oil fraction, up to 10 mL in total (e.g., 7 mL of diethyl ether to 3 mL of the oil). The desired oily mixes were obtained by combining 1 mL of each ether-oil solution and then letting ether evaporate.

3.2.3 CG-MS Analysis

The gas chromatographic/mass spectrometric (GC/MS) EOs analyses were carried out with a GC-MS and GC-FID similarly as previously described[35,100].

3.2.4 Bacterial Strains and Culture Conditions

P. aeruginosa PaO1 was grown in Brain Heart Infusion broth (BHI, Oxoid, Basingstoke, UK). Planktonic cultures were grown in flasks under vigorous agitation (180 rpm) at 37 °C while biofilm formation was assessed in a static condition at 37 °C in 96-well plates for 18 hours.

3.2.5 Determination of Minimal Inhibitory Concentration (MIC)

MIC was performed according to the guidelines of Clinical Laboratory Standards Institute (CLSI). Each EO was added directly from mother stock, and solutions were prepared by two-fold serial dilutions. Mother stock solutions were obtained by solubilizing each EO in DMSO at a final concentration of 1 g/mL. A total of 10 concentrations were used within the 25-0.045 mg/mL range. Experiments were performed in quadruplicate. The MIC was determined as the lowest concentration at which the observable bacterial growth was inhibited. No inhibition of the bacterial growth was highlighted at tested concentrations.

3.2.6 Static Biofilm Assay

Biofilm formation of *P. aeruginosa* PaO1 was evaluated in the presence of each EO. Quantification of in vitro biofilm production was based on previously reported methodology[93]. Briefly, the wells of a sterile 96-well flat-bottomed polystyrene plate were filled with 100 µL of the appropriate medium. A measure of 1/100 dilution of overnight bacterial cultures was added into each well (about 0.5 OD 600 nm). As a control, the first row contained bacteria grown in 100 µL of BHI (untreated bacteria).

Furthermore, BHI broth was added to the remaining wells starting from the third row. In the second row, we added BHI supplemented with each EO at a concentration of 25 mg/mL. Samples were serially diluted (1:2 dilutions) starting from this lane. The plates were incubated aerobically for 18 hours at 37 °C.

Biofilm formation was measured using crystal violet staining. After treatment, planktonic cells were gently removed; each well was washed three times with double-distilled water and patted dry with a piece of paper towel in an inverted position. Each well was stained with 0.1% crystal violet and incubated for 15 minutes at room temperature, rinsed twice with double-distilled water, and thoroughly dried to quantify biofilm formation. The dye bound to adherent cells was solubilized with 20% (v/v) glacial acetic acid and 80% (v/v) ethanol. After 30 minutes of incubation at room temperature, OD590 was measured to quantify the biofilm's total biomass formed in each well. Each data point is composed of four independent experiments, each performed at least in triplicate.

3.2.7 Statistical Analysis of Biological Evaluation

Data reported were statistically validated using Student's *t*-test comparing mean absorbance of treated and untreated samples. The significance of differences between mean absorbance values was calculated using a two-tailed Student's *t*-test. A *p*-value of <0.05 was considered significant.

3.2.8 Machine Learning

3.2.8.1 General methods

Binary classification models development and validation were carried out by an in-house python script based on the scikit-learn machine learning library[72]. First, the data were imported and pre-processed to obtain the independent data matrix consisting of 89 rows (essential oil samples) and 54 columns (chemical components). Two dependent target vectors containing 89 biofilm formation percentage observations at 48 μg/mL and 3.125 mg/mL were defined.

Principal Component Analysis (PCA)[20] was used to check for linear data separability, while Gradient Boosting (GB)[9] for non-linear classification. Cross-validation was used to search for the optimal inhibition percentage cut-off value in order to define active and inactive samples. The optimal cut-off values were used to obtain the final classification model. Hyper-parameter optimization was finally achieved through a systematic grid search of the number of stages to perform (number of trees), maximum depth of individual tree which limits the number of nodes in the tree (max depth), minimum number of samples required to be at a leaf node (min sample leaf), and the number of features to consider (**Table 4**).

Table 4. GB parameters used in the grid search for optimal hyper-parameterization

N estimators	100	250	500	750	1000	1250	1500	1750
Max depth	3	4	5	6	7	8	9	10
Min samples leaf	1	3	5	7	9	11	13	15
Max features	0.7	0.6	0.5	0.4	0.3	0.2	log2[a]	sqrt[b]

[a] sqrt: max features = $\sqrt{(\mathrm{n_{features}})}$

[b] log2: max features = $log2(n_{features})$

The final classification model was numerically and graphically evaluated by accuracy (ACC), Matthews correlation coefficient (MCC), receiver operating characteristic (ROC) (**Figure 8**), and precision-recall (PR) (**Figure 9**) curves. Finally, the importance of EOs chemical components was evaluated individually through the “feature importance” and “partial dependence” plots[9]. Partial dependence plots may be viewed as a graphical representation of linear regression model coefficients that extends to arbitrary model types, addressing a significant component of the model.

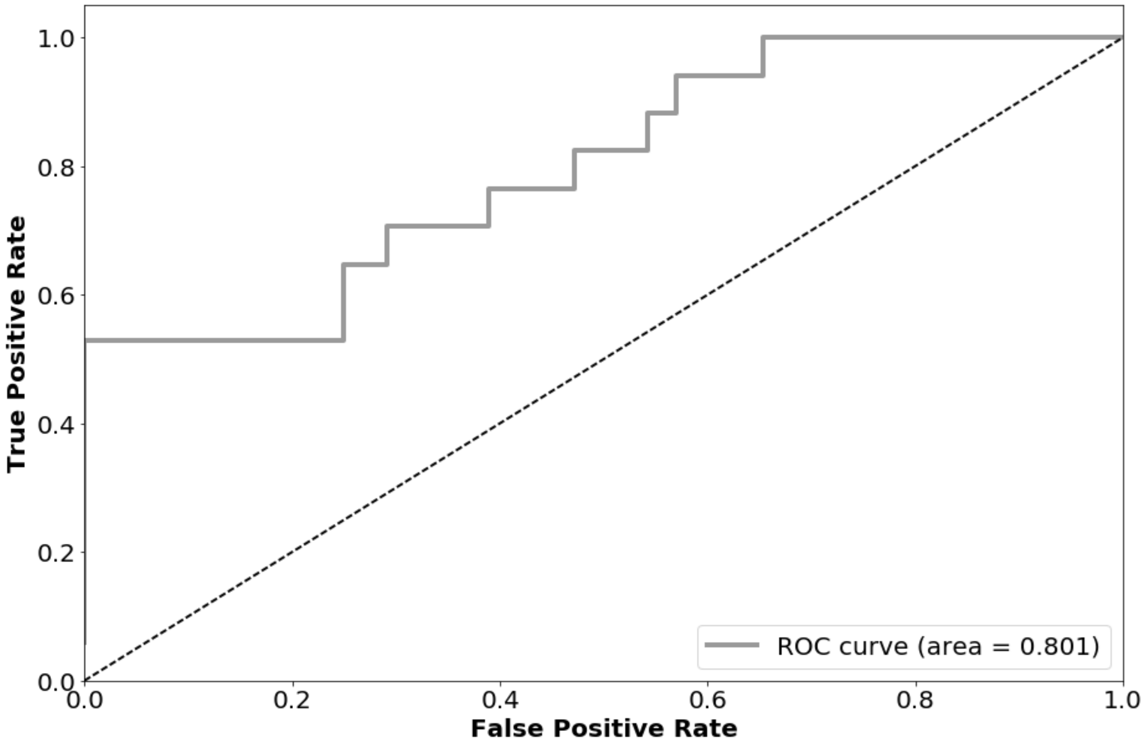


Figure 8. ROC curve for the GB classification model obtained for biofilm inhibition measured at 48.8 µg/ml.

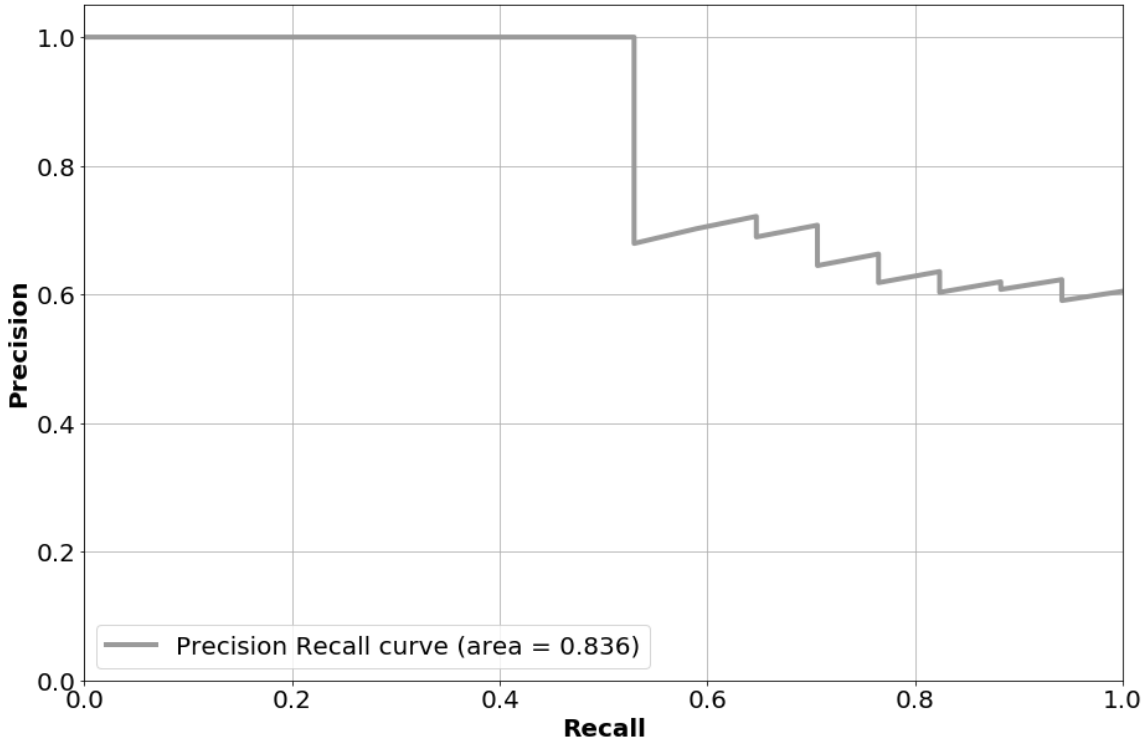


Figure 9. Precision-Recall curve for the GB classification model obtained for biofilm inhibition measured at 48.8 µg/ml.

3.2.8.2 Validation

Validation of the classification model was carried out by leave-one-out cross-validation and taking into account the accuracy (ACC), the precision or positive predictive value (PPV), the recall or sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), receiver operating characteristic (ROC) curve, and the Matthews correlation coefficient (MCC)[101].

3.3 Results

3.3.1 EO Extraction

The fractioned extraction process applied to three different plant species, two of them being also monitored in terms of different harvesting periods, showing great differences in EO yields[36]. In the case of CG, usually, the major parts of EOs were extracted in the first 3 or 6 hours[35]. The great impact of the harvesting period is particularly evident in the case of FVEOs, and a significant increase, of up to five times, in essential oil amount was noticed in October when the plant was fruiting[37]. Annual RS gave a very unusual yield curve with the first maximum after the first hour of extraction and the second one between the third and sixth hours of the extraction process. Relative yield percentages of EOs (calculated per weight of fresh/dried plant material) and total yields over time are given in **Table 5** and

Table *6*.

Table 5. Relative yields % of essential oils over time

Plant species	**h** [1]	**1**	**2**	**3**	**6**	**12**	**24**
	m [2]						
CG	*Jul.*	0.300	0.350	0.360	0.366	0.370	0.373
	Aug.	0.300	0.360	0.400	0.420	0.426	0.432
	Sep.	0.190	0.250	0.300	0.360	0.376	0.381
	Oct.	0.180	0.260	0.290	0.320	0.328	0.328
FV	*Aug.*	0.070	0.110	0.140	0.180	0.196	0.213
	Sep.	0.090	0.140	0.170	0.200	0.218	0.240
	Oct.	0.360	0.640	0.830	1.090	1.210	1.250
RS	*na*	0.200	0.300	0.440	0.640	0.740	0.800

[1] Extraction hour, [2] Month of harvesting, na - not applicable.

Table 6. EO Yield % calculated on the dried (FV and RS) or fresh (CG) plant material.

Plant species	h [1]	1	2	3	6	12	24
	m [2]						
CG	*Jul.*	0.300	0.050	0.010	0.006	0.004	0.003
	Aug.	0.300	0.060	0.040	0.020	0.006	0.006
	Sep.	0.190	0.060	0.050	0.060	0.016	0.005
	Oct.	0.180	0.080	0.030	0.030	0.008	0.0004
FV	*Aug.*	0.070	0.040	0.030	0.040	0.016	0.017
	Sep.	0.090	0.050	0.030	0.030	0.018	0.022
	Oct.	0.360	0.280	0.190	0.260	0.120	0.040
RS	*na*	0.200	0.100	0.140	0.200	0.100	0.060

[1] Extraction hour, [2] Month of harvesting, na - not applicable.

3.3.2 GC-MS Analysis of EOs

Obtained CGEOs, FVEOs, and RSEOs were analyzed in terms of chemical composition[35,37]. The extraction method applied gave fractions that differ significantly in their chemical composition characterized by 89 samples with a total of 54 chemical constituents differently distributed (**Table 7** and **Table 8**). For each EO, the main characterizing compounds are usually present in every fraction, variations in their amount are particularly abundant in the first three fractions (up to 3 hours of extraction process) with a very low percentage, or even absent, in the last three (after 12 or 24 hours). Furthermore, some compounds appear only with the development of the extraction process, being significantly present only in the last fractions. Concerning the harvesting period, essential oil chemical profiles were found to be heavily influenced by this factor. Details for CG and FV (**Table 7**) have already been reported[35,37], while chemical data for RS are reported in table **Table 8**.

Table 7. Chemical composition (%) of the most active FVEO samples.

# [1]	Name	Sample[2]										
		A1h	**A3h**	**A6h**	**A12h**	**AM1**	**AM2**	**AM3**	**AM4**	**S1h**	**OM1**	**OM4**
1	*α*-pinene	-	2.9	4.9	1.1	1.0	7.0	3.9	3.3	1.9	4.3	19.6
2	*β*-pinene	-	-	-	-	-	-	-	-	-	3.3	1.4
3	*β*-terpinene	-	-	-	-	-	0.4	0.9	0.6	-	-	-
4	*β*-myrcene	1.6	1.4	1.5	0.8	0.2	0.4	1.0	1.1	1.4	3.0	2.0
5	*α*-phellandrene	18.3	6.1	20.5	3.9	0.4	0.7	1.0	10.5	13.0	2.7	7.7
6	d-limonene	7.5	7.0	8.2	3.7	1.8	3.4	6.8	6.0	12.4	1.0	4.4
7	*β*-phellandrene	6.6	4.1	4.6	2.2	1.2	2.2	5.0	4.4	6.8	1.1	1.6
8	*γ*-terpinene	1.1	1.1	2.6	0.8	0.1	0.2	1.0	1.1	1.5	3.9	1.6
9	*o*-cymene	35.5	25.2	19.1	14.0	16.0	23.5	35.3	28.7	52.5	1.5	2.9
10	terpinolene	-	-	-	-	-	-	-	-	-	0.1	0.5
11	fenchone	2.4	1.3	0.4	0.2	4.8	2.6	2.0	1.7	-	4.8	8.8
12	dehydro-p-cymene	0.2	0.5	0.6	1.8	-	-	-	-	-	-	-
13	isomenthone	0.5	1.1	1.0	2.2	-	-	-	-	-	-	-
15	pulegone	2.3	4.4	0.3	9.3	4.5	2.2	2.8	3.7	0.4	-	-
16	estragole	18.5	14.5	6.1	1.7	22.5	14.8	14.6	13.2	-	70.6	47.1
17	*p*-ment-en-2-one	-	3.4	3.5	9.7	3.8	1.1	2.0	2.7	0.6	-	-
18	phellandral	-	1.1	2.3	4.1	1.4	1.6	1.1	1.3	0.2	-	-
20	*cis*-sabinol	4.7	8.6	6.3	9.5	7.8	5.6	6.1	6.5	2.2	-	-
21	*p*-cymen-8-ol	-	3.0	1.7	3.5	5.2	2.0	1.8	1.8	0.8	-	1.1
22	2,3-pinanediol	-	-	-	-	7.8	3.3	2.8	5.7	1.3	-	-
26	thymol	-	5.9	9.1	14.7	4.8	6.7	3.7	4.6	1.2	-	-
27	myristicin	-	-	2.2	5.6	1.6	0.8	1.0	-	2.7	1.2	1.1
28	piperitenone oxide	-	-	2.5	5.8	4.9	19.5	4.0	-	-	0.7	-
	Unidentified compounds	**0.8**	**8.4**	**2.6**	**5.4**	**10.2**	**2.0**	**3.2**	**3.1**	**1.1**	**1.8**	**0.2**

[1] # indicates the compound identification number;

[2]samples names were given by merging the month first letter and extraction time as reported in **Table 5** or by merging the first letter of the month, the letter M (mixture) and serial number of the mix.

Table 8. Chemical composition (%) of the most active RSEO samples.

# [1]	**Name**	**Sample** [2]			
		1h	**3h**	**12h**	**30h**
1	*α*-pinene	3.9	3.6	1.2	-
2	*β*-pinene	4.6	3.4	1.5	0.1
3	*β*-myrcene	0.9	1.2	0.4	-
4	*α*-phellandrene	-	-	5.4	0.6
5	*d*-limonene	7.4	6.7	1.2	0.3
6	*β*-terpinene	3.0	4.9	1.8	0.3
7	*β*-ocimene	0.5	1.3	0.8	0.2
8	*o*-cymene	40.1	3.8	7.4	4.2
9	terpinolene	-	2.1	1.6	-
11	borneol	-	3.3	3.0	3.0
12	pulegone	-	-	0.8	2.6
13	citral	-	1.0	-	-
14	cryptone	-	2.5	-	-
15	*p*-menth-1-en-2-one	-	2.2	2.1	9.7
17	*cis*-sabinol	-	5.8	4.3	12.9
18	*p*-cymen-8-ol	9.2	13.4	3.0	6.4
19	piperitenone oxide	6.5	3.6	1.0	1.9
21	2,3-pinanediol	9.6	1.8	2.1	-
23	myristicin	-	-	3.2	1.7
24	apiol	6.5	21.1	59.2	56,1
	Unidentified compounds	**7.8**	**18.3**	**0.0**	**0.0**

[1] # indicates the compound identification number;

[2]samples names indicate the extraction time, as reported in **Table 5**.

3.3.3 Qualitative Analysis of EOs Effect on Biofilm Formation of *P. aeruginosa*

In order to exclude if selected EOs contained molecules affecting bacterial viability, the 89 EOs were also analyzed for antimicrobial activity. In vitro EOs bacteriostatic and bactericidal activities were evaluated on *P. aeruginosa* by broth microdilution methods. An appropriate dilution (10^6 cfu/mL were used as reported by National Committee for Clinical Laboratory Standards NCCLS, 2004) of the bacterial culture of *P. aeruginosa* in the exponential phase was used. No antimicrobial activity on *P. aeruginosa* strains was highlighted for all tested EOs (maximal concentration tested 25 mg/mL). The anti-biofilm effects of EOs from different plants above described were examined on *P. aeruginosa* PAO1. Firstly, we selected some representative EOs samples (two for RS, three for CG, and FV) to evaluate the anti-biofilm efficacy at different concentrations starting from 25 mg/mL using scalar dilutions (**Table 9**).

Table 9. Effect of EOs at different concentrations (scalar concentrations starting from 25 mg/mL) on biofilm formation of *P. aeruginosa* PaO1. Data are reported as the percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of four independent experiments, each performed at least in triplicate.

EO (mg/mL)	R3	R12	CJM3	CAM4	CSM2	FS1	FSM5	FOM4
25	55.11	50.62	36.71	59.48	28.23	30.84	47.38	28.31
12.5	41.41	45.18	37.10	54.56	41.36	38.83	49.12	25.48
6.25	37.77	57.44	34.64	55.82	37.79	30.16	49.51	25.01
3.125	42.25	57.42	40.09	71.80	40.48	38.40	54.16	25.62
1.55	48.49	65.06	38.80	69.33	44.34	44.93	90.16	30.44
0.78	47.81	64.60	50.05	67.19	51.65	38.67	78.74	37.15
0.39	49.49	61.97	54.87	72.45	42.97	84.39	76.10	39.34
0.18	57.39	66.48	53.90	69.42	49.18	60.02	80.81	32.54
0.09	60.37	61.83	48.00	72.80	43.08	59.75	78.51	38.32
0.0488	70.65	59.05	52.99	83.24	50.26	42.44	88.71	38.28
0.0244	45.12	63.91	41.65	73.93	34.01	47.96	59.63	39.47
0.0122	64.81	66.11	46.59	73.19	40.02	57.26	75.63	38.29
0.0061	65.40	59.87	50.14	82.00	37.86	27.50	143.53	37.34
0.00305	63.06	78.37	45.22	69.05	35.44	38.70	117.45	39.75
0.00152	60.94	70.11	44.76	79.77	40.72	44.94	104.92	47.53
0.00076	61.95	65.18	40.29	73.17	37.14	37.13	112.35	53.00
0.0003814	61.13	62.05	49.74	76.76	47.15	42.07	113.75	37.23
0.0001907	56.98	65.80	48.48	83.21	49.49	36.59	90.67	57.15
0.00009535	72.29	65.27	45.52	71.44	52.18	39.25	79.33	46.41
0.000047675	64.71	74.79	44.19	91.78	46.23	43.30	99.52	68.74

The obtained preliminary data were analyzed in terms of biofilm reduction and reproducibility, which led to selecting two concentrations as the most representatives (3.125 mg/mL and 0.0488 mg/mL). The first concentration was in the range of milligrams. The second one was in the range of micrograms. The percentages of residual biofilm after treatment at these two concentrations (3.125 mg/mL and 0.0488 mg/mL) for all 89 essential oils are reported in **Figure 10**. It is worth noticing that each EO had a specific effect on biofilm formation, thus depending on its characteristic and unique composition, which was previously quali-quantitatively analyzed chemically. Furthermore, the majority of tested EOs had an inhibitory effect on *P. aeruginosa* PaO1 biofilm formation. Arbitrarily, three biofilm inhibition levels were considered for clustering the EOs potencies qualitatively: potent biofilm inhibition in the range 0–40% of residual biofilm, mild inhibition in the range 40-80%, and no biofilm inhibition over 80% of residual biofilm, respectively.

In some cases, an increase in biofilm formation was highlighted after the treatment. As reported in **Figure 10A**, almost all EOs samples derived from FV showed to be able to inhibit biofilm formation of *P. aeruginosa* PaO1. The only exception was the FO1 sample. A marked effect dose-dependent was observable (i.e., FA2, FSM5, FO3, FO6, and FOM4, where the anti-biofilm effect was proportional to the concentration of EO used). In **Figure 10B**, the effects of EOs from CG on PaO1 biofilm formation are reported. Differently from FV data, several CGEOs samples showed increasing biofilm production directly proportional to the concentration used. Among all assayed EOs, some of them, such as CO2, produced an inhibitory effect on the biofilm at higher concentrations and increased it at lower concentrations. Instead, other EOs can strongly inhibit biofilm formation already at very low concentrations (reduction of biofilm higher than 50%). Regarding the results obtained with extracts derived from RS, all of them inhibited biofilm. In most cases, the reduction is proportional to the concentration of EO used (R1, R3, R6, R24, R30, RM2, RM3, and RM4). Conversely, EOs named R2, R12, RM5, and RM6 did not show an anti-biofilm effect correlated to the concentration used. Only in the case of RM1, there is an opposite relationship between the concentration used and the anti-biofilm effect (**Figure 10C**).

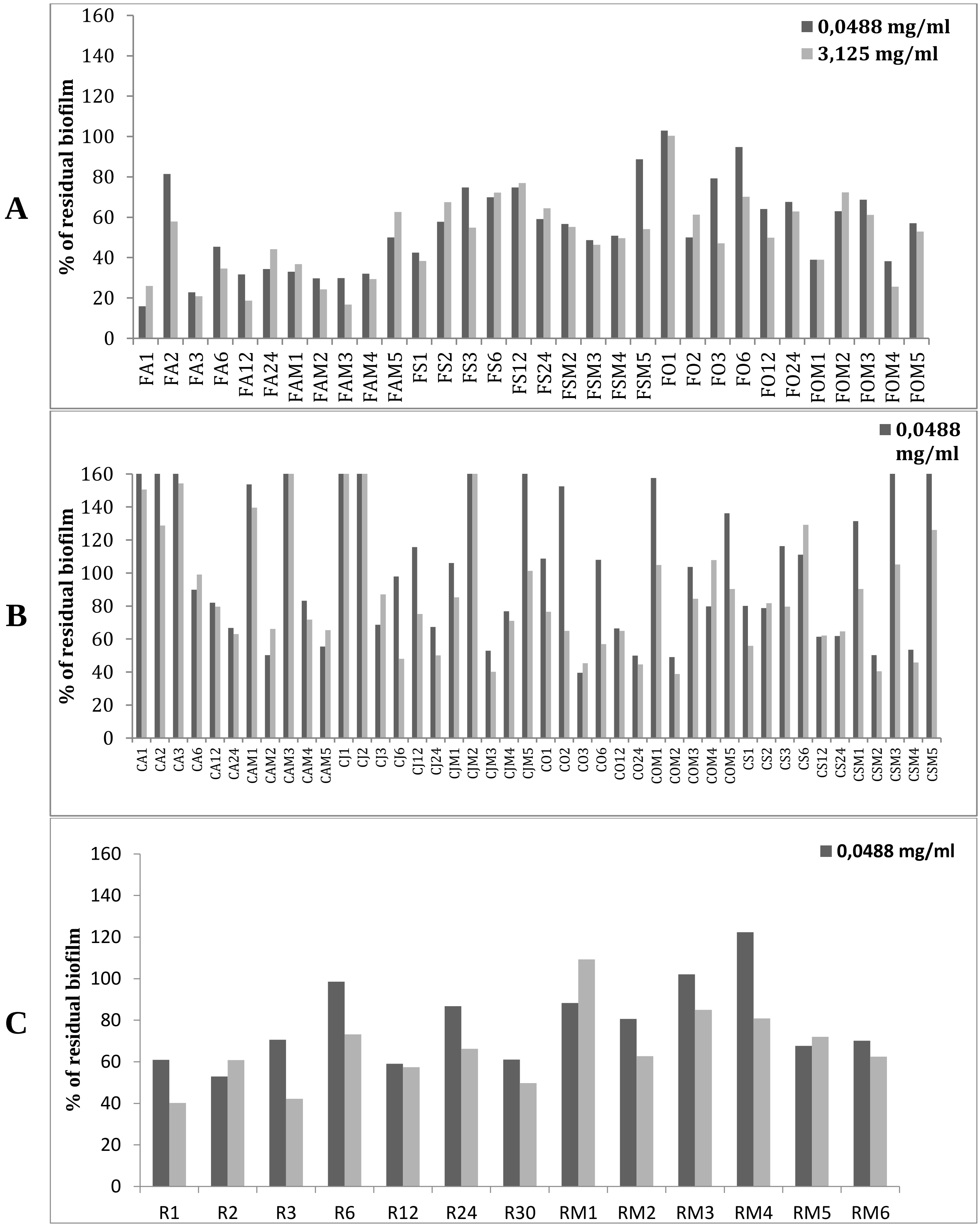


Figure 10. Effect of EOs from Foeniculum vulgare Miller (FV) (**A**), Calamintha nepeta (L.) Savi subsp. glandulosa (Req.) Ball (CG) (**B**), and Ridolfia segetum Moris (RS) (**C**) on biofilm formation of *P. aeruginosa* PaO1. Data are reported as percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of four independent experiments each performed at least in triplicate.

Table 10 summarizes the results of the anti-biofilm activity of EOs grouped in four different classes corresponding to their capability to impair biofilm formation. This classification was based on results reported in **Figure 10** obtained with a higher concentration of EOs (3.125 mg/mL). In particular, a substantial biofilm reduction was judged if the residual biofilm was in the range 0-40%, medium reduction if the residual biofilm was in the range 40-80%, and no reduction for residual biofilm higher than 80%. In such cases, an enhancer effect on biofilm formation was evidenced after the treatment with EOs. This latter was observable only for such EOs obtained from CG.

Table 10. An arbitrary classification of 89 EOs samples in 4 different classes depending on their capability to impair biofilm formation.

Strong reduction (< 40% residual biofilm)			**Medium reduction (40-80% residual biofilm)**			**No reduction (80 - 100% residual biofilm)**			**Enhancer effect (>100% residual biofilm)**		
FV	**CG**	**RS**	**FV**	**CG**	**RS**	**FV**	**CG**	**RS**	**FV**	**CG**	**RS**
FA1	COM2		FA2	CJM3	R1	FO1	CA6	RM1		CA1	
FA3			FA24	CJM4	R2		CJ3	RM3		CA2	
FA6			FAM5	CO1	R3		CJM1	RM4		CA3	
FA12			FS2	CO2	R6		CJM5			CAM1	
FAM1			FS3	CO3	R12		COM1			CAM3	
FAM2			FS6	CO6	R24		COM3			CJ1	
FAM3			FS12	CO12	R30		COM5			CJ2	
FAM4			FS24	CO24	RM2		CS2			CJM2	
FS1			FSM2	CS1	RM5		CS3			COM4	
FOM1			FSM3	CS12	RM6		CSM1			CS6	
FOM4			FSM4	CS24						CSM3	
			FSM5	CSM2						CSM5	
			FO2	CSM4							
			FO3	CA12							
			FO6	CA24							
			FO12	CAM2							
			FO24	CAM4							
			FOM2	CAM5							
			FOM3	CJ6							
			FOM5	CJ12							
				CJ24							

3.3.4 Application of Machine Learning Algorithms

Initial application of linear models by the PCA formalism revealed the lack of linear dependence between biofilm production inhibition and chemical composition as no acceptable classification model was obtained using biofilm production percentages observed at either 0.0488 or 3.125 mg/mL. Indeed, a visual inspection of the scores plot of the first two principal components (PCs), accounting for more than 75% of data variance, revealed the presence of at least three clusters (**Figure 11**).

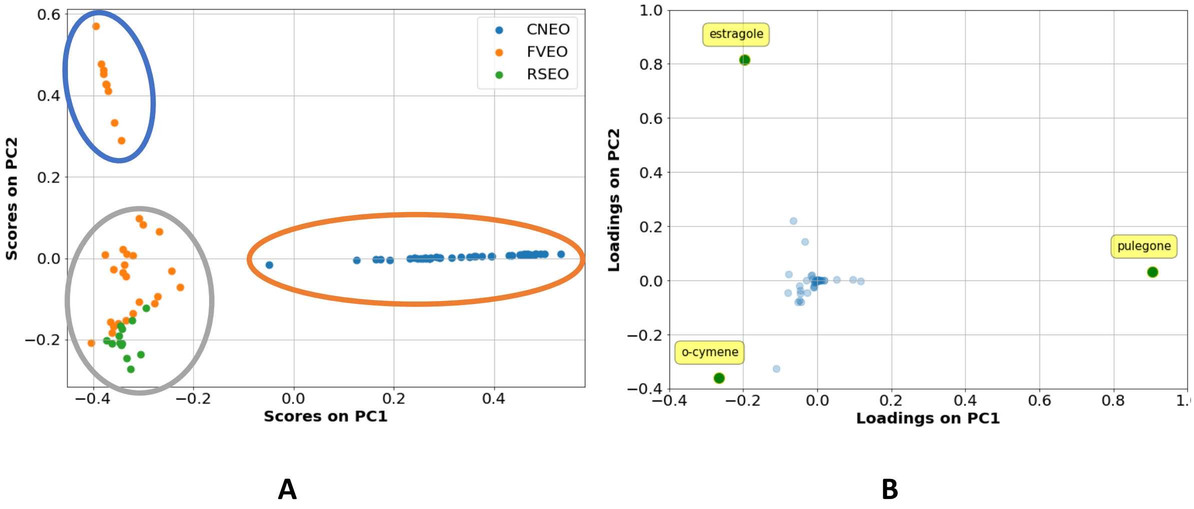
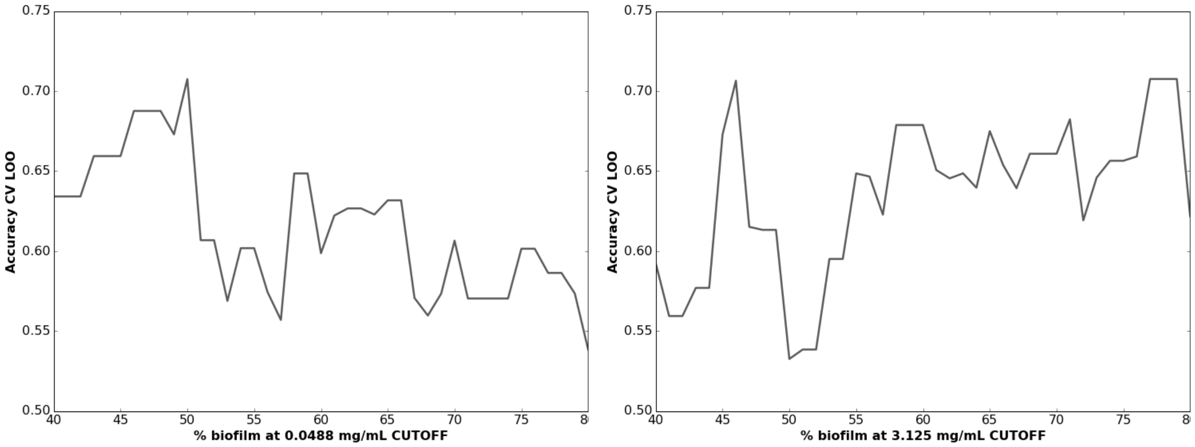


Figure 11. PCA first 2 PCs graphical plots. The core plot (**A**) indicates the presence of at least three clusters (circled in (**A**)). The loading plots (**B**) highlights that estragole, o-cymene, and pulegone could be the most important chemical constituents among all the tested EOs.

Concurrently, from **Figure 11**, PCA identified the three plants' derived essential oils, although it was impossible to obtain a clear separation between Ridolfia and Foeniculum EOs. PCA-related loading plots indicated that estragole, o-cymene, and pulegone chemical components were the most important chemical constituents among all the tested EOs. Furthermore, the scores plot (**Figure 11A**) highlighted the lack of any linear classification among the 89 EOs as the three clusters cannot be associated with any level of biofilm production percentage. Linear classification models using algorithms such as logistic regression (LR) and linear support vector machines[7] did not lead to satisfying classifiers. Therefore, non-linear algorithms like

Random Forest (RF)[8], non-linear Support Vector Machine (SVM)[102], and Gradient Boosting (GB)[9] were applied.

Among the used algorithms, GB led to the most robust binary classification model. To this aim, at first, the optimal biofilm production percentage for the binary classification was investigated by systematically increasing it from a starting 40 to 80% and monitoring the accuracy by leave-one-out cross-validation. The best GB classification model was obtained at 50% and 46% for the 48.8 μg/mL and 3.125 mg/mL concentrations, respectively (**Figure 12**). Therefore, EOs characterized by more than 50% (or 46%) of biofilm production were classified as inactive. Those with lower values were considered active.



A **B**

Figure 12. Analysis of best cutoff values for the GB classification models at 48.8 μg/ml (**A**) and at 3.125 mg/ml (**B**).

The two models were characterized by good statistical values (**Table 11**). In particular, greater robustness was obtained for the classification model defined at 48.8 μg/mL oil concentration as highlighted from ACC, MCV, precision-recall AUC, and ROC AUC higher values (**Figure 8** and **Figure 9**).

Table 11. Cross-validation scores for the binary GB classification models [a].

Statistical Parameter	At 48.8 µg/mL	At 3.125 mg/mL
ACC CV	0.90	0.72
MCC CV	0.64	0.51
Precision–Recall AUC	0.84	0.72
ROC AUC	0.80	0.68

[a] final optimized models were obtained with the following settings: max depth = 3, max features = 0.9, min samples_leaf = 16, n estimators = 500.

3.4 Discussion

This study aimed to address the potential of selected EOs to prevent and treat biofilm produced by *P. aeruginosa*. This microorganism is widespread and is a frequent foodborne pathogen. Although *P. aeruginosa* is an opportunistic pathogen and rarely causes disease in healthy persons, it is a notorious nosocomial pathogen, posing a high risk to immunosuppressed individuals and other highly vulnerable populations patients[103]. *P. aeruginosa* can cause pneumonia, catheter-associated and urinary tract infections, and sepsis in wounded patients, sometimes resulting in chronic severe infections and health complications. The ability of *P. aeruginosa* to form biofilm renders it refractory to the action of antibiotics and disinfectants and able to survive in unfavorable conditions for a long time. Based on these considerations, it is evident the importance of having new strategies to impair biofilm formation by *P. aeruginosa*.

3.4.1 Chemical Quantitative Composition-Activity Relationships

Significant to moderate biofilm reducing activity was observed for several CGEO samples against *P. aeruginosa* PaO1 strain. Analyzing data showed in **Figure 10**, the extraction process's duration seems to influence the activity on this strain, since in every month (except October), the last fractions (12 and 24 hours) were found to be more effective. The observed efficacy of these last fractions could be potentially associated with the increase of chrysanthenone. Some FVEO samples from the August harvest demonstrated remarkably high biofilm inhibition of this strain, in some samples, even more than 80%. With few exceptions, September samples did not show any significant reduction, while several FVEOs obtained from fruiting material (FOM1 and FOM4) showed significant ability to impair biofilm formation, even in the relatively low concentration.

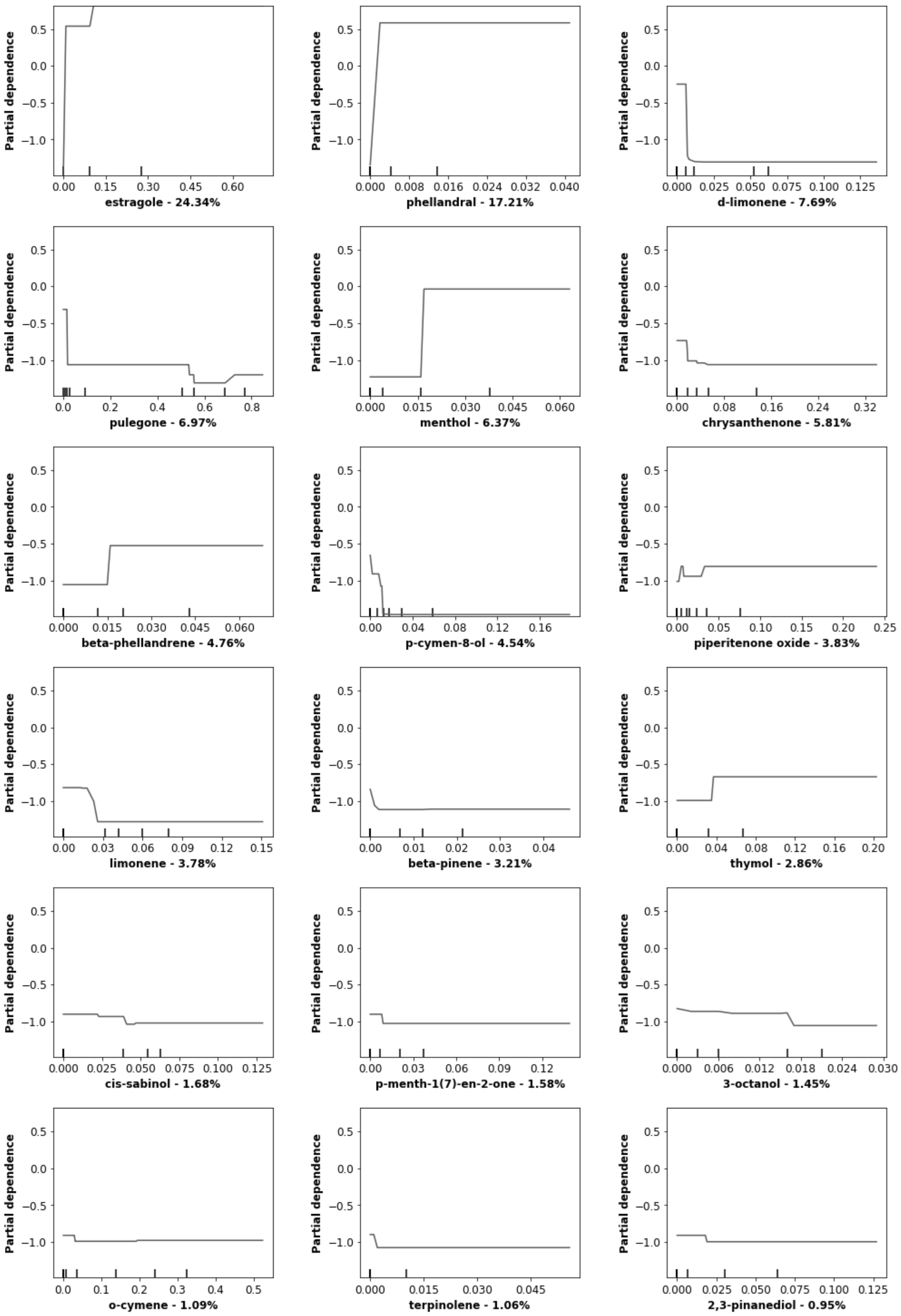
Further data analysis (**Figure 11**) suggests the potential influence of different chemical profiles on oil's efficacy. Namely, October samples differ from the ones obtained during the pre-fruiting phenological stage, having estragole as the principal constituent. The presence of this phenylpropene in these samples may be the main reason for biofilm inhibition. On the other side, the higher susceptibility of *P. aeruginosa* strain to the oils obtained from August harvesting may be influenced by some other components characteristic of the EOs obtained in that period that possibly exerts some additive effect in the expression of overall activity (α-phellandrene, β-phellandrene, or thymol).

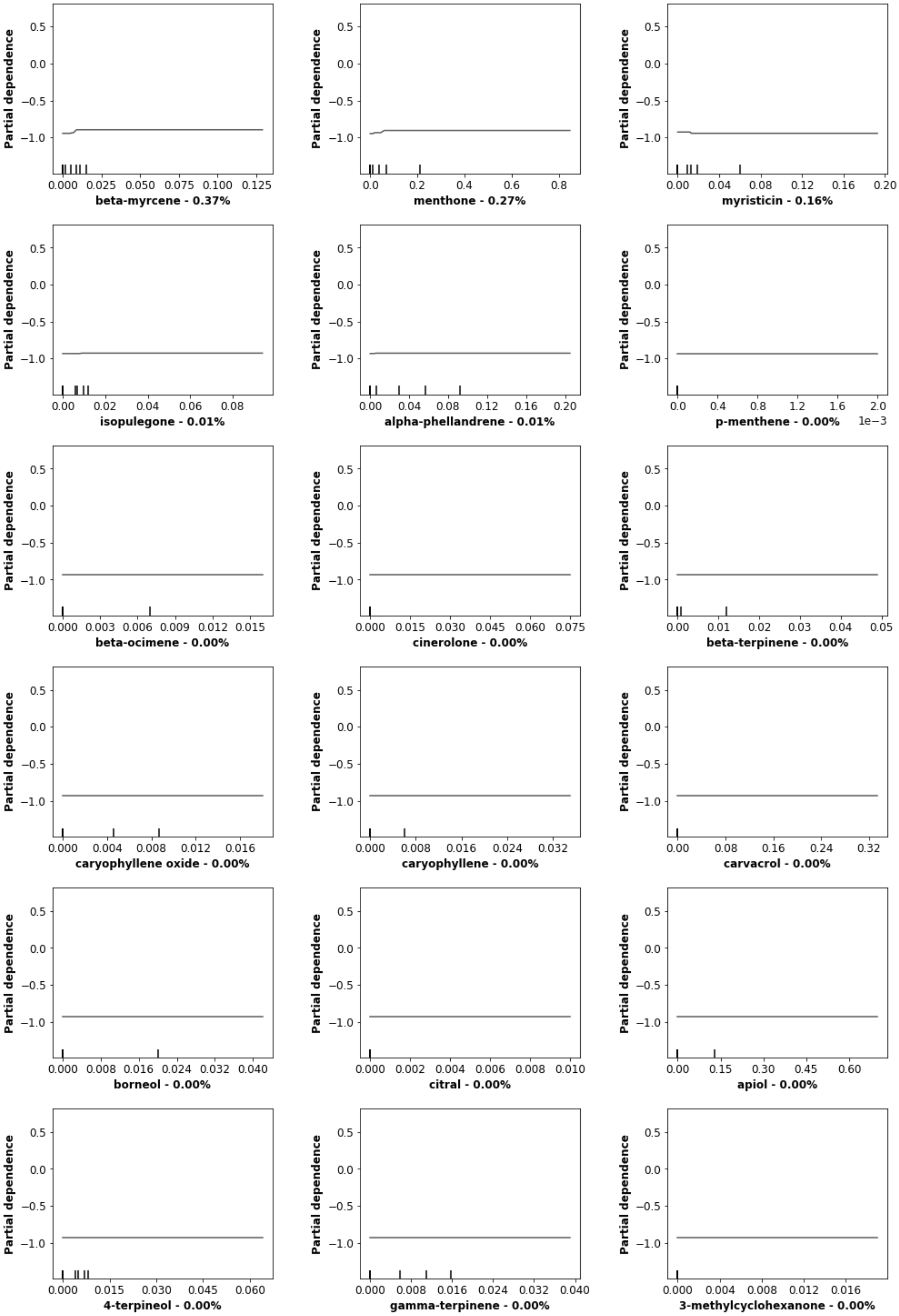
P. aeruginosa PaO1 was sensitive only to specific RSEO fractions (**Figure 10C**). However, taking into account the chemical analysis of the samples, a positive correlation cannot be established between the content of o-cymene and apiol, the main characterizing compounds, and the inhibition of biofilm formation. Probably some minor components may influence biofilm inhibition.

In such cases in the presence of EOs obtained by CG, an enhancer effect on biofilm formation was observed. This second effect is an exciting result and supports the theory that plants produce molecules that regulate biofilm formation in different environments: a fascinating example of inter-kingdom regulation. The regulatory pathways of the sessile phenotype could be related to the competitive dynamics of habitats. Identifying the molecules responsible for these mechanisms could be attractive in opening new perspectives for the control of bacterial biofilm formation. It is worth noting that the EOs that we used represent a complex pool of chemical cues that could be characterized by different capabilities to either impair or promote biofilm formation.

3.4.2 Gradient Boosting Binary Classification Model

The most robust classification model defined at 48.8 µg/mL oil concentration was analyzed through partial dependence (**Figure 13**) and feature importance (**Figure 14**) plots[9]


Partial dependence
0.5
0.0
−0.5
−1.0
0.00 0.15 0.30 0.45 0.60
estragole - 24.34%
0.000 0.008 0.016 0.024 0.032 0.040
phellandral - 17.21%
0.000 0.025 0.050 0.075 0.100 0.125
d-limonene - 7.69%
0.0 0.2 0.4 0.6 0.8
pulegone - 6.97%
0.000 0.015 0.030 0.045 0.060
menthol - 6.37%
0.00 0.08 0.16 0.24 0.32
chrysanthenone - 5.81%
0.000 0.015 0.030 0.045 0.060
beta-phellandrene - 4.76%
0.00 0.04 0.08 0.12 0.16
p-cymen-8-ol - 4.54%
0.00 0.05 0.10 0.15 0.20 0.25
piperitenone oxide - 3.83%
0.00 0.03 0.06 0.09 0.12 0.15
limonene - 3.78%
0.00 0.01 0.02 0.03 0.04
beta-pinene - 3.21%
0.00 0.04 0.08 0.12 0.16 0.20
thymol - 2.86%
0.000 0.025 0.050 0.075 0.100 0.125
cis-sabinol - 1.68%
0.00 0.03 0.06 0.09 0.12
p-menth-1(7)-en-2-one - 1.58%
0.000 0.006 0.012 0.018 0.024 0.030
3-octanol - 1.45%
0.0 0.1 0.2 0.3 0.4 0.5
o-cymene - 1.09%
0.000 0.015 0.030 0.045
terpinolene - 1.06%
0.000 0.025 0.050 0.075 0.100 0.125
2,3-pinanediol - 0.95%


Partial dependence
0.5
0.0
−0.5
−1.0
0.000 0.025 0.050 0.075 0.100 0.125
beta-myrcene - 0.37%
0.0 0.2 0.4 0.6 0.8
menthone - 0.27%
0.00 0.04 0.08 0.12 0.16 0.20
myristicin - 0.16%
0.00 0.02 0.04 0.06 0.08
isopulegone - 0.01%
0.00 0.04 0.08 0.12 0.16 0.20
alpha-phellandrene - 0.01%
0.0 0.4 0.8 1.2 1.6 2.0
1e−3
p-menthene - 0.00%
0.000 0.003 0.006 0.009 0.012 0.015
beta-ocimene - 0.00%
0.000 0.015 0.030 0.045 0.060 0.075
cinerolone - 0.00%
0.00 0.01 0.02 0.03 0.04 0.05
beta-terpinene - 0.00%
0.000 0.004 0.008 0.012 0.016
caryophyllene oxide - 0.00%
0.000 0.008 0.016 0.024 0.032
caryophyllene - 0.00%
0.00 0.08 0.16 0.24 0.32
carvacrol - 0.00%
0.000 0.008 0.016 0.024 0.032 0.040
borneol - 0.00%
0.000 0.002 0.004 0.006 0.008 0.010
citral - 0.00%
0.00 0.15 0.30 0.45 0.60
apiol - 0.00%
0.000 0.015 0.030 0.045 0.060
4-terpineol - 0.00%
0.000 0.008 0.016 0.024 0.032 0.040
gamma-terpinene - 0.00%
0.000 0.004 0.008 0.012 0.016
3-methylcyclohexanone - 0.00%

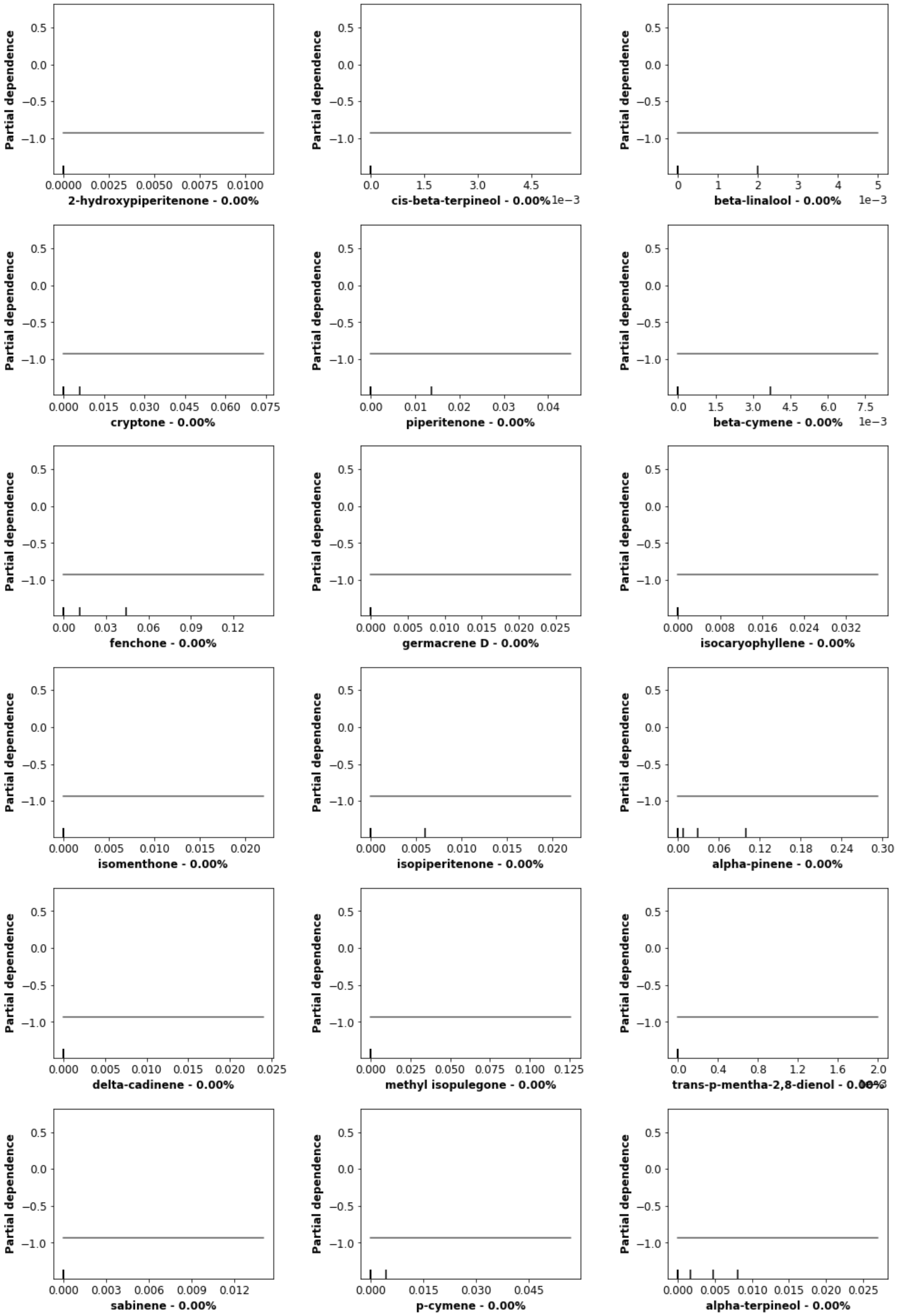


Figure 13. Partial Dependence plot obtained for the GB classification models at 48.8 µg/ml for the EOs' chemical components

Feature importance plot highlights each chemical constituent's absolute importance, while partial dependence plots built for the most important components give direct univariate relationships with the biofilm inhibitory activity, giving direct information on positive or negative effects.

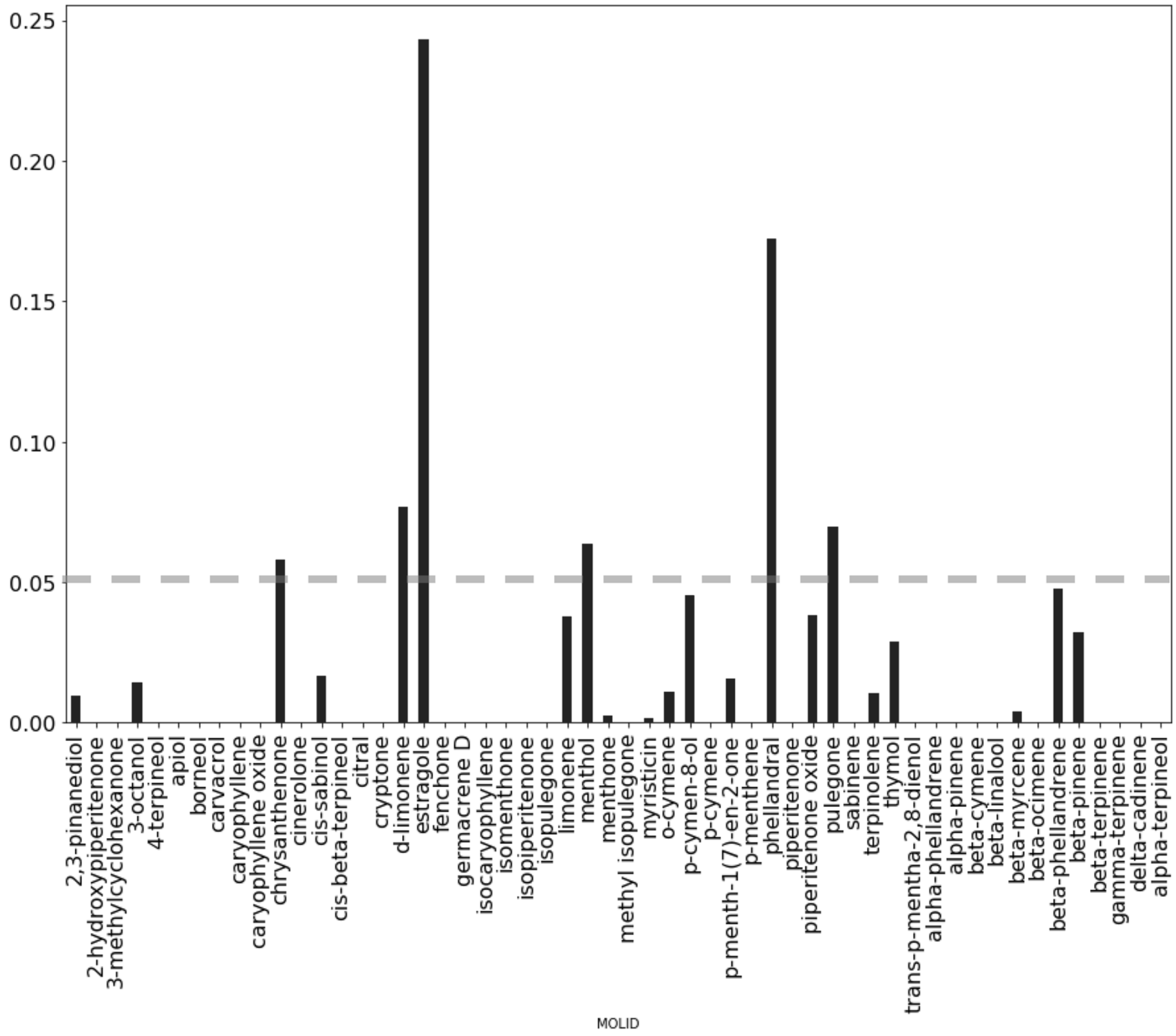


Figure 14. Feature importance plot obtained for the GB classification models at 48.8 µg/mL.

From **Figure 14**, considering a threshold of 5%, six compounds, namely estragole, methol and phellandral, d-limonene, pulegone, and chrysanthenone, can be considered as those most influencing biofilm productions, being estragole and phellandral the most significant. The partial dependence plots were investigated for all compounds (**Figure 13**) to ascertain whether components correlate positively or negatively with biofilm inhibition. The partial dependence plots for the above six compounds (**Figure 15**) directly indicate estragole and phellandral as those most critical for biofilm inhibition.

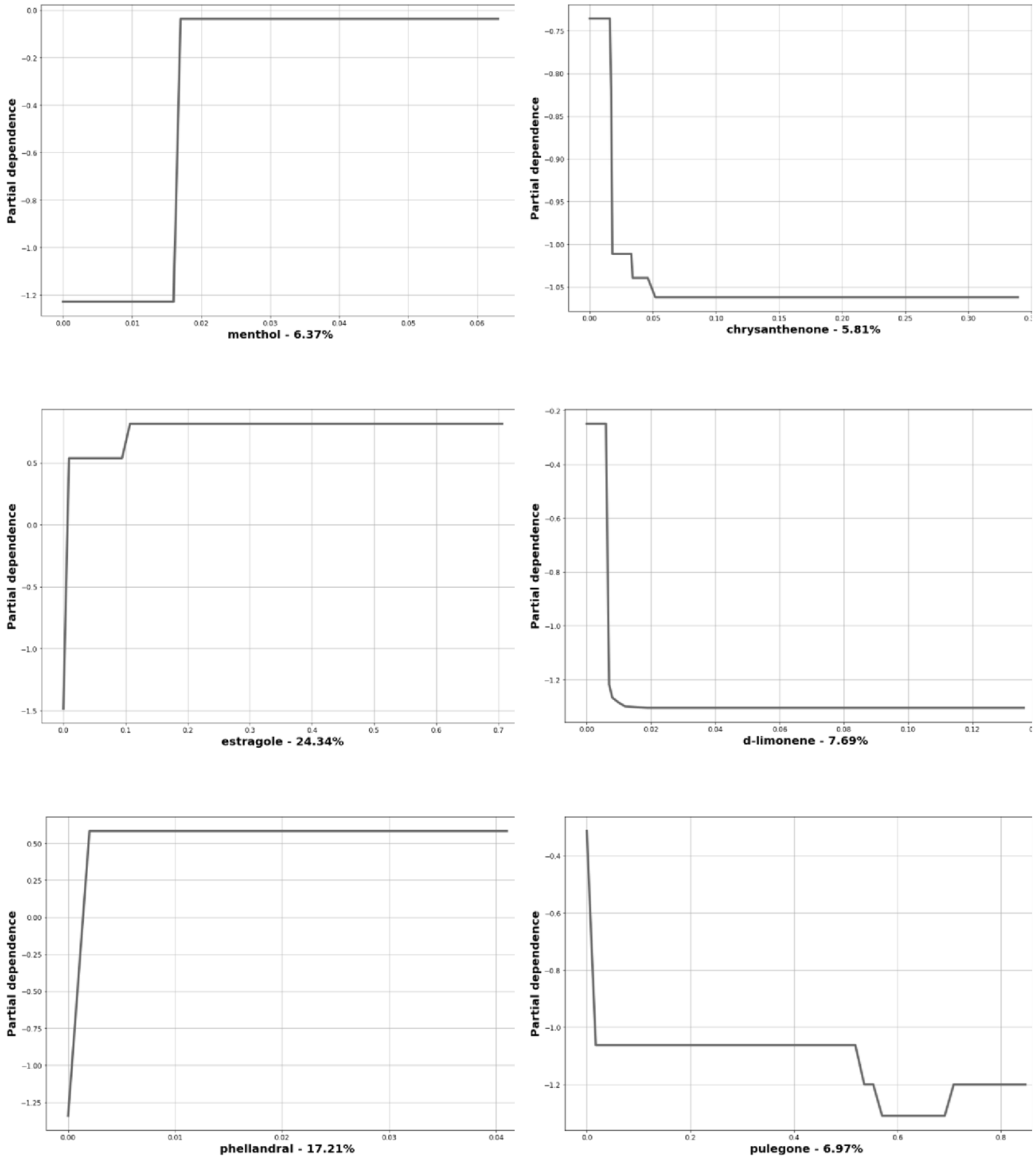


Figure 15. Partial Dependence plot obtained for the Gradient Boosting (GB) classification models at 48.8 µg/ml for the most important chemical components.

Whereas d-limonene, pulegone, and chrysanthenone were inversely associated with biofilm inhibition and were likely responsible for biofilm production enhancement observed. A different scenario can be interpreted for menthol; at a low percentage, it correlates with a negative effect on biofilm inhibition that disappears as it is increased above 16%.

3.5 Conclusions

This study demonstrated that the biofilm growth of *P. aeruginosa* PAO1 is influenced by the presence of EOs extracted from three different Mediterranean plants harvested in different seasons. These results suggest that the kind of biofilm modulation depends on EO chemical composition, although the fractions were obtained from the same plant. Remarkably, a significant influence on the modulation of biofilm production is related to the harvesting period. Furthermore, in some cases, the same EOs seem to exert opposite influences (stimulation or inhibition of biofilm growth) depending on sample dilution. This is related to the concentration of specific chemical compounds, as highlighted by the classification models. The biofilm change in growth, in the presence of the essential oils, is possibly due to a modulation of the phenotype that switches from biofilm to planktonic since the action is not related to a bacteriostatic/bactericidal activity on *P. aeruginosa*. The latter could be explained with the presence in EOs of small molecules, likely acting as quorum sensing inhibitors. In any case, moderation on conclusion has to be undertaken since interactive and synergistic effects among the EO chemical components, including minor ones, can affect biological potency. The application of an *ad hoc* developed python-based machine learning protocol led to the definition of a classification model able to discriminate essential oils in active and inactive at a cut-off value of 50% of biofilm formation using a concentration of 48.8 µg/mL. Investigation of the most critical components through feature importance and partial dependence plots seems to indicate estragole and phellandral as the chemical components mostly related to biofilm inhibition, while d-limonene, pulegone, and chrysanthenone seem to be related to biofilm production. As validated by five performance metrics, the classification model is an example showing machine learning as a tool to investigate complex chemical mixtures, and possibly in future experiments. It could enable scientists to understand the mechanism by which EOs act. Based on these results, further experiments are on due course to investigate EOs rich in the above five chemical components to validate the classification model. As this research's primary goal was focused on the evaluation of antimicrobial and antibiofilm potencies and a vast number of different essential oils, no investigation was undertaken on EO effects on mature biofilms and their eradication. The data from this study, enriched by further experiments carried out with other EOs and bacterial species, could enable the identification of blends of EOs specifically designed to obtain products with strong anti-biofilm efficacy applicable in many fields airborne decontamination, products for dermatological and respiratory tract infections.

4 Machine Learning Analyses on Data including Essential Oil Chemical Composition and In Vitro Experimental Antibiofilm Activities against *Staphylococcus* Species

4.1 Introduction

A biofilm is a microbially derived sessile community characterized by cells irreversibly attached to a substrate or interface or each other, embedded in a self-produced matrix of extracellular polymeric substances, which exhibits an altered phenotype concerning growth, gene expression, and protein production[104]. Biofilm resistance to antimicrobials[105] is a complex phenomenon, driven not only by genetic mutation-induced resistance but also via increased microbial cell density that supports resistance through horizontal gene transfer across cells[106]. Indeed, other mechanisms are involved, such as (i) low penetration of antimicrobial agents due to the barrier function exerted by the biofilm matrix, (ii) presence of cells exhibiting a high multidrug tolerance, (iii) reduced susceptibility to antibiotics as a consequence of stress adaptive responses or changes in the chemical biofilm microenvironment[107]. The strategies adopted to treat these challenging infections are rapidly changing due to the increasing understanding of biofilm structure and functions. Nonetheless, the prevention of biofilm formation and the treatment of existing biofilms is currently a complex challenge; therefore, the discovery of new multi-targeted or combinatorial therapies is increasingly urgent[108]. Therefore, the development of anti-biofilm agents is considered of significant interest and represents an important strategy since non-biocidal molecules to avoid the rapid appearance of resistant mutants are highly valuable. Staphylococci are prevalent causes of biofilm-associated infections among bacteria[109]. In particular, *Staphylococcus aureus* (*S. aureus*) is an opportunistic pathogen that can cause severe diseases in humans, ranging from skin and soft tissue infections to invasive infections of the bloodstream, heart, lungs, and other organs[110]. In 2013, Nicholson et al. reported that 30% of the U.S. population was colonized by *S. aureus,* while 1.5% was found to be a carrier of methicillin-resistant *S. aureus* (MRSA), a major cause of healthcare-related infections responsible for a significant proportion of nosocomial infections worldwide. Recently in the

U.S., deaths from MRSA infections have exceeded those from many other infectious diseases, including HIV/AIDS[111]. *Staphylococcus epidermidis* (*S. epidermidis*), conventionally considered a commensal of human skin, can cause significant problems when breaching the epithelial barrier, especially during biofilm-associated infection of indwelling medical devices[112,113]. Most diseases caused by *S. epidermidis* exhibit a chronic profile and occur as device-related infections (such as an intravascular catheter or prosthetic joint infections) and/or their complications[113].

Given the above scenario, the scientific community seeks new agents endowed with anti-biofilm capabilities to fight *S. aureus* and *S. epidermidis* infections. Recently, several reports indicated in vitro efficacy of non-biocidal essential oils (EOs) as a promising treatment to reduce bacterial biofilm production and prevent the inducing of drug resistance[42]. In different applications, EOs have been found of some efficacy in reducing biofilm production of either *S. aureus* standard strains or MRSA[114–119]. In other reports, EOs and some of their purified chemical components have also been proved to inhibit *S. epidermidis* biofilm production[120–122].

Machine learning (ML) has been proved as a tool able to profoundly investigate the modulatory role of EOs' chemical components on *Pseudomonas aeruginosa* biofilm production[43,123–125]. In particular, 89 EOs extracted in different periods and times of extractions from three different plants were analyzed: 13 EOs (RSEOs) from *Ridolfia segetum* Moris (RS); 32 EOs (FVEOs) from *Foeniculum vulgare* Miller (FV) and 44 EOs (CGEOs) from *Calamintha nepeta* (L.) Savi subsp. *glandulosa* (Req.) Ball, (CG)[43]. In line with that study and to investigate EOs' ability also to reduce bacterial biofilm production in other bacteria, herein is reported an extensive study of the 89 EOs samples as potential antibacterial and anti-biofilm agents against *S. aureus* ATCC 6538P, *S. aureus* ATCC 25923, *S. epidermidis* RP62A and *S. epidermidis* O-47. To this purpose, ML algorithms were applied to the EOs' chemical compositions and the determined associated anti-biofilm potencies, to shed light on those components likely mainly responsible for either positive or negative modulation of biofilm production.

4.2 Materials and Methods

4.2.1 Essential oil and Chemical Composition Analysis

Essential oils and their chemical compositions were available from previously reported studies[35,37,43]. Essential oils were obtained by direct fractionated steam distillation and analyzed by gas chromatographic/mass spectrometric (GC/MS) protocol[36,100].

4.2.2 Bacterial Strains and Culture Conditions

Bacterial strains used in this work (**Table 12**) were grown in Brain Heart Infusion broth (BHI, Oxoid, UK). Biofilm formation was assessed in static conditions. Planktonic cultures were grown in flasks under vigorous agitation (180 rpm) at 37 °C. In particular, *S. aureus* ATCC 6538P (6538P) and *S. aureus* ATCC 25923 (25923) are reference strains for antimicrobial testing; *S. epidermidis* RP62A (RP62A) is a reference strain isolated from an infected catheter, while *S. epidermidis* O-47 (O-47) is a clinical isolate intense biofilm producer strain characterized by a genomic mutation in agr locus[126].

Table 12. Details of the used bacterial strains.

Strain	Name	Type	Isolation
S. aureus 6538P	6538P	clinical isolate	ATCC collection
S. aureus 25923	25923	clinical isolate	ATCC collection
S. epidermidis RP62A	RP62A	infected catheter isolated strain	ATCC collection
S. epidermidis O-47	O-47	septic arthritis clinical isolate	Heilmann *et al.*, 1996

4.2.3 Determination of Minimal Inhibitory Concentration (MIC)

The MIC was determined as the lowest concentration at which the observable bacterial growth was inhibited. MICs were determined according to the guidelines of Clinical Laboratory Standards Institute[127] (CLSI). Each EO was added directly from mother stock, and two-fold serial dilutions prepared solutions. Mother stock solutions were obtained by solubilizing each EO in DMSO at a final concentration of 1 g/mL. Appropriate dilution (10^6 cfu/mL) of bacterial culture in the exponential phase was used. Ten concentrations were used within the 25-0.045 mg/mL range. Experiments were performed in quadruplicate.

4.2.4 Biofilm Production Assay

The quantification of biofilm production was based on microtiter plate biofilm assay (MTP): a suitable dilution of bacterial culture in the exponential growth phase was added into wells of a sterile 96-well flat-bottomed polystyrene plate in the absence and in the presence of each EO. Quantification of in vitro biofilm production was based on previously reported methodology[93]. The wells of a sterile 96-well flat-bottomed polystyrene plate were filled with 100 µL of the appropriate medium. 1/100 dilution of overnight bacterial cultures were added into each well (about 0.5 OD 600nm). As a control, the first row contained bacteria grown in 100 µL of BHI (untreated bacteria). In the second row was added BHI supplemented with each EO at concentrations of 3.125 mg/mL and 0.0488 mg/mL, respectively. The plates were incubated aerobically for 18 hours at 37 °C. Biofilm formation was measured using crystal violet staining. After treatment, planktonic cells were gently removed; each well was washed three times with double-distilled water and patted dry with a piece of paper towel in an inverted position. Each well was stained with 0.1% crystal violet and incubated for 15 minutes at room temperature, rinsed twice with double-distilled water, and thoroughly dried to quantify biofilm formation. The dye bound to adherent cells was solubilized with 20% (v/v) glacial acetic acid and 80% (v/v) ethanol. After 30 min of incubation at room temperature, OD590 was measured to quantify the biofilm's total biomass formed in each well. Each data point is composed of 4 independent experiments, each performed at least in 3-replicates. EOs altering biofilm formation of selected strains were then tested, as reported below. Briefly, the wells of a sterile 96-well flat-bottomed polystyrene plate were filled with 100 µL of the appropriate medium. 1/100 dilution of overnight bacterial cultures were added into each well (about 0.5 OD 600nm). The first row contained bacteria grown in 100 µL of BHI (untreated bacteria) as a control.

Furthermore, BHI broth was added to the remaining wells starting from the third row. In the second row was added BHI supplemented with each EO at a concentration of 0.0488 mg/mL. Samples were diluted serially (1:2 dilutions) starting from this lane. The plates were incubated aerobically for 18 hours at 37 °C. Biofilm formation was measured using crystal violet staining, as previously reported.

4.2.5 Statistical Analysis of Biological Evaluation

Data reported were statistically validated using Student's *t*-test comparing mean absorbance of treated and untreated samples. The significance of differences between mean absorbance values was calculated using a two-tailed Student's *t*-test. A p-value of <0.05 was considered significant.

4.2.6 Machine Learning Binary Classification

4.2.6.1 General Methods

All calculations were performed using the Python (version 3.6, https://www.python.org/) programming language[128] by executing in-house code in the Jupyter Notebook platform (version 5.7)[129]. The datasets were imported and loaded into a Pandas[70,71] data frame and pre-processed to obtain four independent data matrices consisting of 89 rows (essential oil samples) and 54 columns (chemical components). Two dependent target vectors containing 89 biofilm production percentage observations at 48 µg/mL and 3.125 mg/mL were defined. Machine learning algorithms used in this study were implemented using the sklearn library[72] (version 0.20). Unsupervised dimensionality reduction was performed with Principal component analysis[20] (PCA), while L2 regularized logistic regression was used for the supervised learning analysis. The scores and loadings relatives to the first two principal components (PCs) were graphically inspected on plots generated using the matplotlib[69] library (version 3.0). Thirty principal components were extracted for each dataset to build the classification models. Cross-validation was used to search for the optimal inhibition/activation percentage cut-off values in order to define active and inactive samples. The optimal cut-off values were used to obtain the hyper-parameters optimized classification models. Hyper-parameters optimization was achieved through a Bayesian optimization[15] of the number of PCs to be used as features and the regularization parameter of the L2-Logistic Regression (inverse of regularization strength in the sklearn implementation). For each dependent target vector, two types of models were built: one to define EOs' ability to inhibit biofilm production and another to describe biofilm production enhancement. Percentage ranges of 60-80% and 120-140% biofilm productions were chosen for inhibition and activation models, respectively. Finally, the most appropriate cut-offs for binary classification of biofilm inhibitors/not-inhibitors or biofilm enhancers/not-enhancers EOs were determined from a supervised learning analysis.

The binary classification models were numerically and graphically evaluated by accuracy (ACC), Matthews correlation coefficient (MCC), receiver operating characteristic (ROC), and precision-recall (PR) curves. Finally, EOs chemical components' importance was evaluated individually through the feature importance and partial dependence plots[9] as implemented in the Skater python library[130,131]. Feature importance is a generic term for the degree to which a predictive model relies on a particular feature. Skater feature importance implementation is based on an information-theoretic criterion, measuring the entropy in the change of predictions, given a perturbation of a given feature.

4.2.6.2 Classification Models' Validation

Validation of each classification model was carried out by leave-one-out cross-validation and taking into account the accuracy (ACC), the precision or positive predictive value (PPV), the recall or sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), receiver operating characteristic (ROC) curve and the Matthews correlation coefficient[101] (MCC) Y-scrambling[12,132] was ultimately applied to check any lack of chance correlation and assess coefficients robustness.

4.3 Results

4.3.1 Antimicrobials Activity of EOs

Some FVEOs and CGEOs samples showed MICs at the highest used concentration (**Table 13** and **Table 14**). No antimicrobial activity on staphylococci was recorded for any of the RSEOs, except for the R30 sample that showed a MIC value of 25 mg/mL (**Table 15**). Only seven out of 89 EOs displayed MIC values of 6.25 mg/mL against the two *S. aureus* strains (**Table 13**, **Table 14**, and **Table 15**). As a control, MIC was also evaluated for ofloxacin, a conventional antibiotic belonging to the fluoroquinolone family.

Table 13. MIC determined for CGEOs on *Staphylococcus* spp strains. EOID indicates sample names. Sample names are the same as previously reported[43]. Ofloxacin MIC is also reported as a positive reference drug. All data are expressed in mg/mL.

EOID	*S. aureus* 6538P	*S. aureus* 25923	*S. epidermidis* RP62A	*S. epidermidis* O-47
CJ1	>25	>25	>25	>25
CJ2	>25	>25	50	>25
CJ3	25	25	>25	>25
CJ6	25	25	>25	>25
CJ12	25	>25	>25	>25
CJ24	>25	>25	>25	>25
CJM1	12.5	25	>25	>25
CJM2	>25	>25	>25	>25
CJM3	12.5	25	25-12.5	25-12.5
CJM4	12.5	25	25	>25
CJM5	>25	>25	>25	>25
CA1	>25	>25	>25	>25
CA2	>25	>25	>25	>25
CA3	25	>25	>25	>25
CA6	25-12.5	>25	25	25
CA12	12.5	12.5	12.5	>25
CA24	12.5	12.5	>25	>25
CAM1	>25	>25	>25	>25
CAM2	12.5-6.25	12.5-6.25	12.5	12.5
CAM3	>25	>25	>25	>25
CAM4	6.25	6.25	12.5	12.5
CAM5	>25	>25	>25	>25
CS1	>25	>25	>25	>25
CS2	>25	>25	>25	>25
CS3	>25	>25	>25	>25
CS6	>25	>25	>25	>25
CS12	12.5-6.25	12.5	12.5	12.5
CS24	6.25	12.5-6.25	>25	12.5
CSM1	>25	>25	>25	>25
CSM2	12.5	12.5	12.5	12.5
CSM3	>25	>25	>25	>25
CSM4	12.5-6.25	12.5-6.25	12.5	12.5
CSM5	>25	>25	>25	>25

Table 13. MIC determined for CGEOs on *Staphylococcus* spp strains. EOID indicates sample names. Sample names are the same as previously reported[43]. Ofloxacin MIC is also reported as a positive reference drug. All data are expressed in mg/mL.

EOID	*S. aureus* 6538P	*S. aureus* 25923	*S. epidermidis* RP62A	*S. epidermidis* O-47
CO1	>25	>25	>25	>25
CO2	>25	>25	>25	>25
CO3	25-12.5	25	25	25
CO6	25-12.5	25-12.5	25-12.5	25-12.5
CO12	12.5	12.5	12.5	12.5
CO24	>25	>25	>25	>25
COM1	25	25	25-12.5	25-12.5
COM2	25-12.5	25-12.5	25-12.5	25-12.5
COM3	25	25	25-12.5	25-12.5
COM4	12.5	25	25	25
COM5	>25	>25	>25	>25
Ofloxacin	0.0002-0.0004	0.0004-0.0008	0.0002-0.0004	0.0002-0.0004

Table 14. MIC determined for FVEOs samples on *Staphylococcus* spp strains. EOID indicates sample names. Sample names are the same as previously reported[43]. Ofloxacin MIC is also reported as a positive reference drug. All data are expressed in mg/mL.

EOID	*S. aureus* 6538P	*S. aureus* 25923	*S. epidermidis* RP62A	*S. epidermidis* O-47
FA1	>25	>25	>25	>25
FA2	25-12.5	25-12.5	>25	>25
FA3	>25	>25	>25	>25
FA6	>25	>25	>25	>25
FA12	>25	>25	>25	>25
FA24	12.5-6.25	>25	>25	>25
FAM1	>25	>25	>25	>25
FAM2	>25	>25	>25	>25
FAM3	>25	>25	>25	>25
FAM4	>25	>25	>25	>25
FAM5	>25	>25	>25	25
FS1	>25	>25	>25	>25
FS2	>25	>25	>25	>25
FS3	>25	>25	>25	>25

Table 14. MIC determined for FVEOs samples on *Staphylococcus* spp strains. EOID indicates sample names. Sample names are the same as previously reported[43]. Ofloxacin MIC is also reported as a positive reference drug. All data are expressed in mg/mL.

EOID	***S. aureus* 6538P**	***S. aureus* 25923**	***S. epidermidis* RP62A**	***S. epidermidis* O-47**
FS6	>25	>25	>25	>25
FS12	>25	>25	>25	>25
FS24	>25	>25	>25	>25
FSM1	>25	>25	>25	>25
FSM2	>25	>25	>25	>25
FSM3	12.5-6.25	>25	>25	>25
FSM4	>25	>25	>25	>25
FSM5	25-12.5	12.5	>25	>25
FO1	25	25	>25	>25
FO2	>25	>25	>25	>25
FO3	25	25	>25	>25
FO6	25	25	>25	>25
FO12	>25	>25	>25	>25
FO24	>25	>25	>25	>25
FOM1	>25	>25	>25	>25
FOM2	>25	>25	>25	>25
FOM3	25-12.5	12.5	>25	>25
FOM4	>25	>25	>25	>25
FOM5	>25	>25	>25	>25
Ofloxacin	0.0002-0.0004	0.0004-0.0008	0.0002-0.0004	0.0002-0.0004

Table 15. MIC determined for RSEOs samples on *Staphylococcus* spp strains. EOID indicates sample names. Sample names are the same as previously reported[43]. Ofloxacin MIC is also reported as a positive reference drug. All data are expressed in mg/mL.

EOID	*S. aureus* 6538P	*S. aureus* 25923	*S. epidermidis* RP62A	*S. epidermidis* O-47
R1	>25	>25	>25	>25
R2	>25	>25	>25	>25
R3	>25	>25	>25	>25
R6	>25	>25	>25	>25
R12	>25	>25	>25	>25
R24	25	>25	>25	>25
R30	25	25	25	25
RM1	>25	>25	>25	>25
RM2	>25	>25	>25	>25
RM3	>25	>25	>25	>25
RM4	>25	>25	>25	>25
RM5	>25	>25	>25	>25
RM6	>25	>25	>25	>25
Ofloxacin	0.0002-0.0004	0.0004-0.0008	0.0002-0.0004	0.0002-0.0004

4.3.2 Biofilm Production Modulation by EOs at Selected Fixed Concentrations

Preliminarily, the same representative EOs (2 RSEOs, 3 CGEOs, and 3 FVEOs) among the reported 89 used on *P. aeruginosa*[43] were selected to evaluate the anti-biofilm potency at different concentrations starting from 25 mg/mL, using scalar dilutions.

The obtained preliminary data analyzed in terms of biofilm production modulation and reproducibility led to the selection of two representatives concentrations (3.125 mg/mL and 0.0488 mg/mL). The first concentration was in the range of milligrams, while the second one was in the range of micrograms. All 89 EOs were then tested at the two selected concentrations, and the biofilm production was measured relative to untreated bacteria (**Figure 16**, **Figure 17**, and **Figure 18**).

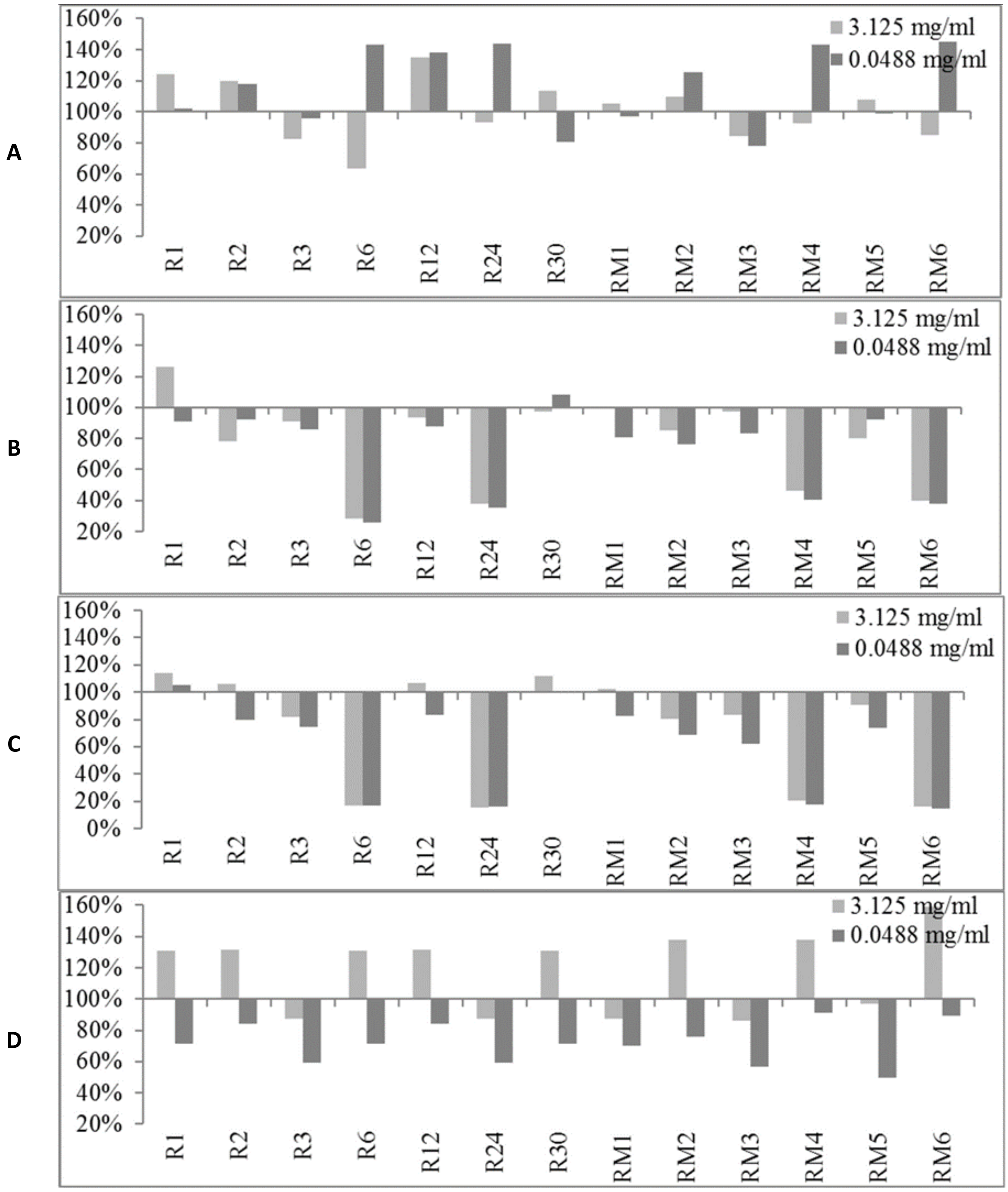


Figure 16. Percentages of biofilm production after treatment at two concentrations (3.125 mg/mL and 0.0488 mg/mL) for RSEOs against the four strains *S. aureus* 6538P (**A**) and 25923 (**B**), *S. epidermidis* RP62A (**C**) and O-47 (**D**, respectively). In the ordinate axis are reported the percentage of bacterial biofilm production. Data are reported as percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments each performed with at least three replicates.

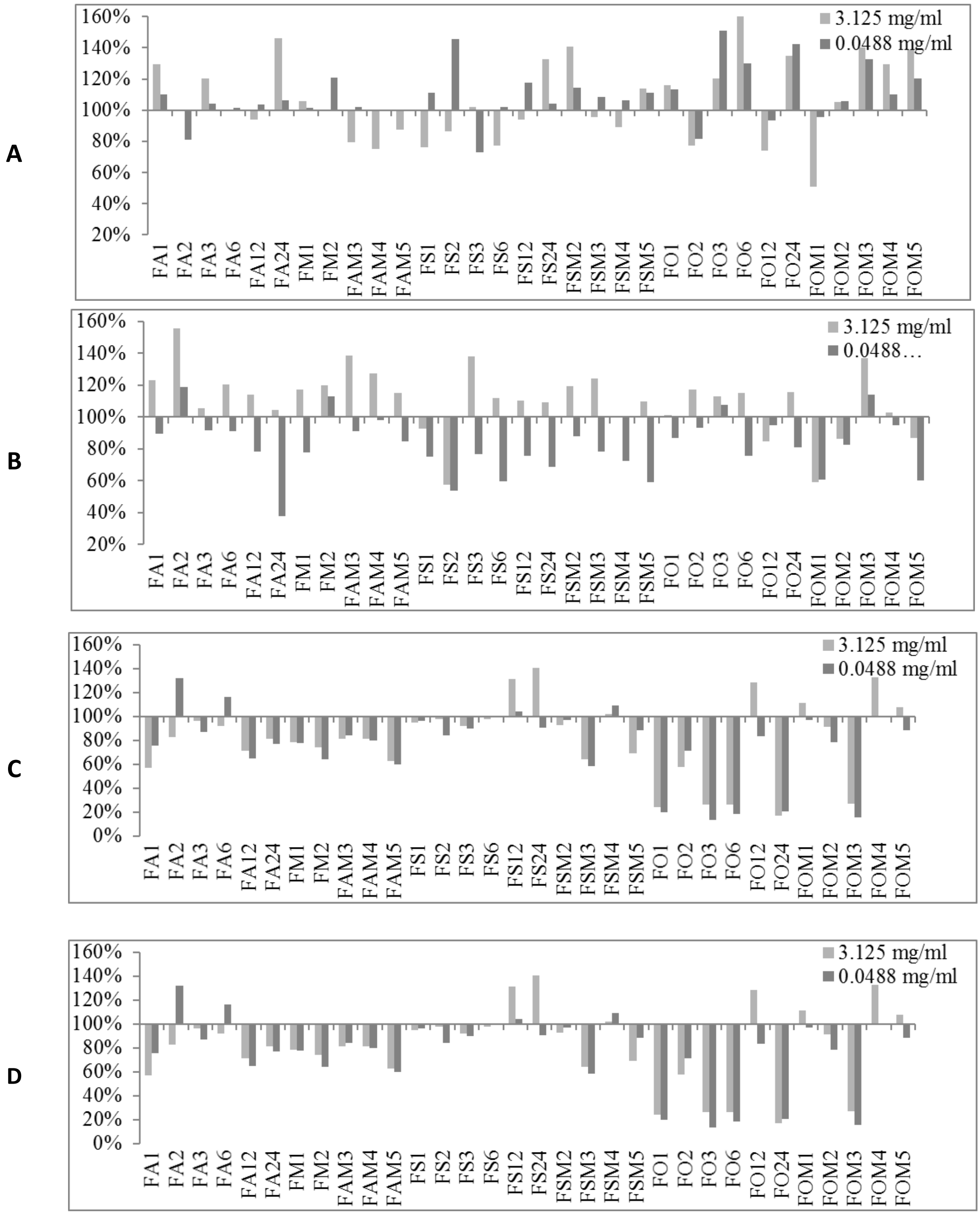


Figure 17. Percentages of biofilm production after treatment at two concentrations (3.125 mg/mL and 0.0488 mg/mL) for FVEOs against the four strains *S. aureus* 6538P (**A**) and 25923 (**B**), *S. epidermidis* RP62A (**C**) and O-47 (**D**, respectively). In the ordinate axis are is reported the percentage of bacterial biofilm production. Data are reported as the percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments, each performed with at least three replicates.

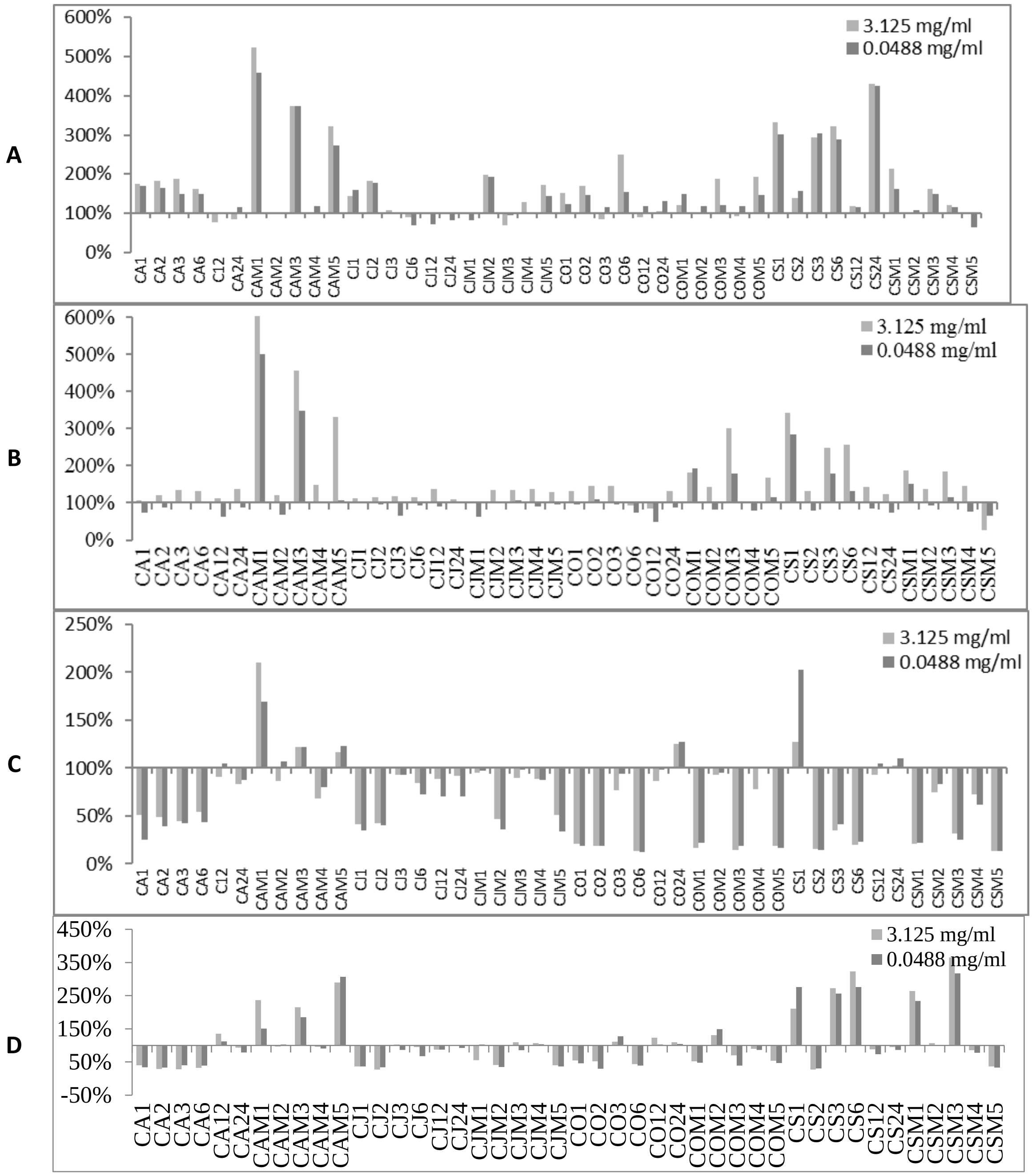


Figure 18. Percentages of biofilm production after treatment at two concentrations (3.125 mg/mL and 0.0488 mg/mL) for CGEOs against the four strains *S. aureus* 6538P (**A**) and 25923 (**B**), *S. epidermidis* RP62A (**C**) and O-47 (**D**), respectively). In the ordinate axis are is reported the percentage of bacterial biofilm production. The abscissa axis is centered at 100% biofilm production. Data are reported as the percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments, each performed with at least three replicates.

At either selected concentration, EOs modulated the biofilm production with unpredictable results for each strain. These results anticipated that many EOs might act mainly as biofilm inhibitors in the case of RP62A and O-47 strains, while for 6538P and 25923, EOs can either induce no effect or stimulate biofilm production (**Table 16**). In **Table 16**, the number of EOs able to inhibit (<100%, <80%, and <50%, respectively) or stimulate (≥100%, ≥120%, ≥150%, and ≥200%, respectively) biofilm formation is reported. It is worthy to note that on *S. epidermidis* strains, about 30 EOs inhibited more than 50% of biofilm growth even at the lowest concentration, while almost none of them showed activity on *S. aureus* strains.

Table 16. Data analysis of biofilm production modulation by EOs at the two selected concentrations as reported in **Figure 16**, **Figure 17** and **Figure 18**.

Conc. µg/mL	*S.* spp Strains	Biofilm Production %		Number EOs Samples at Biofilm Production %						
		MIN	MAX	<50%	<80%	<100%	≥100%	≥120%	≥150%	≥200%
3125	6538P	50.98	523.83	0	10	31	58	38	22	9
	25923	26.92	697.45	1	4	16	73	47	14	9
	RP62A	13.04	209.69	26	42	71	18	8	1	1
	O-47	27.12	289.88	24	35	61	28	14	4	4
48.8	6538P	62.80	459.46	0	5	20	69	37	17	7
	25923	37.91	501.01	3	34	67	22	10	7	3
	RP62A	11.79	202.57	29	48	74	15	6	2	1
	O-47	0.44	306.60	31	48	74	15	7	4	2

4.3.3 Quantitative Analysis of Selected EOs against Different Strains of *S. epidermidis*

Representative EOs selected among those able to reduce more than 70% of biofilm formation were further analyzed to evaluate a dose-dependent effect against *S. epidermidis* RP62A and O-47 (**Figure 19**, **Figure 20**, and **Figure 21**). The inhibition by RSEOs was confirmed at lower concentrations on both strains despite their different biofilm matrix composition, and the inhibition of biofilm formation was clearly not dose-dependent (**Figure 19**). Analogous results were obtained with FVEOs samples (**Figure 20**). Differently, CGEOs revealed a dose-dependent biofilm inhibition being more pronounced on the strongest biofilm producer *S. epidermidis* O-47 than on *S. epidermidis* RP62A (**Figure 21**).

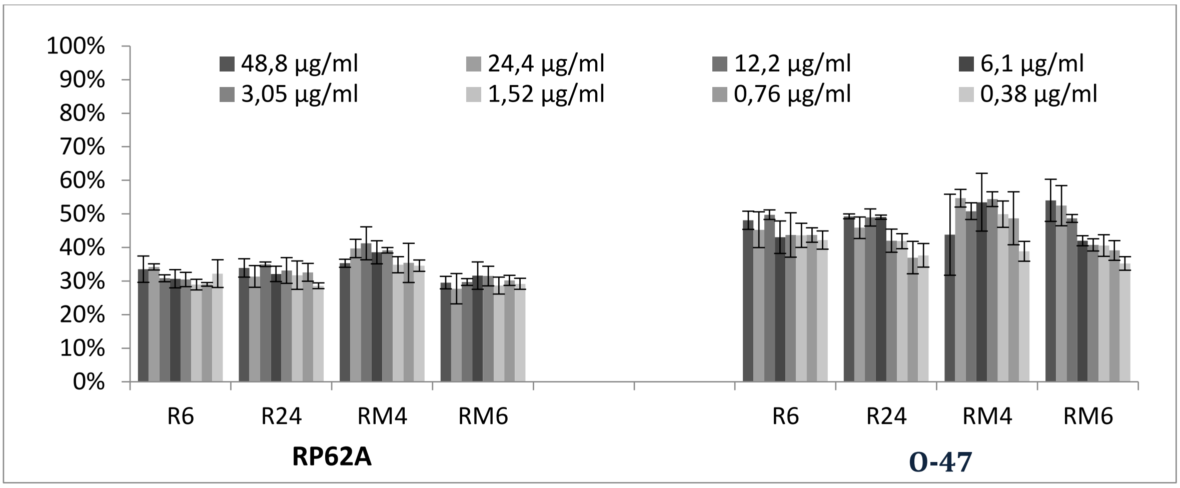


Figure 19. Antibiofilm effect of selected RSEOs on RP62A and on O-47 strains. In the ordinate axis is reported the percentage of bacterial biofilm production. Data are reported as percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments each performed with at least in three replicates.

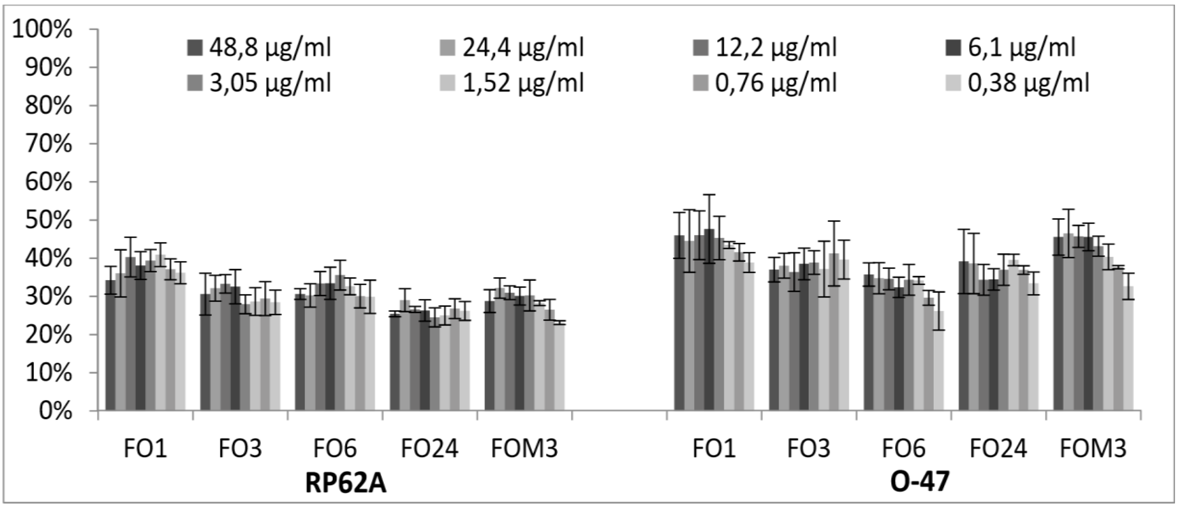


Figure 20. Antibiofilm effect of selected FVEOs on RP62A and on O-47 strains. In the ordinate axis is reported the percentage of bacterial biofilm production. Data are reported as the percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments, each performed with at least three replicates.

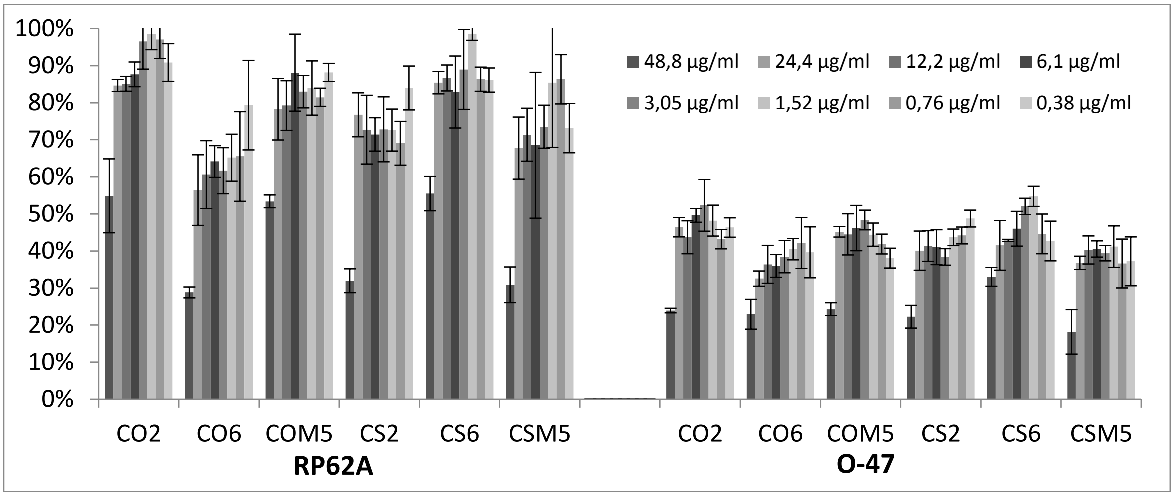


Figure 21. Antibiofilm effect of selected CGEOs on RP62A and on O-47 strains. In the ordinate axis is reported the percentage of bacterial biofilm production. Data are reported as the percentage of residual biofilm after the treatment in comparison with the untreated one. Each data point is composed of 4 independent experiments, each performed with at least three replicates.

4.3.4 Application of Machine Learning Algorithms

4.3.5 PCA Analysis of Datasets

Extraction of the first 3 PCs afforded to a cumulative explained variance of almost 90% (PC1:61.18%; PC1 + PC2: 75.02%; PC1 + PC2 + PC3: 82.92%). The first two principal components indicate at least three clusters (**Figure 22**) and correctly identified the three plants derived EOs, although no definite separation between RSEOs and FVEOs was observable. PCA related loading plots indicated estragole, o-cymene, and pulegone as the chemical components mainly related to high PCs values (**Figure 22**).

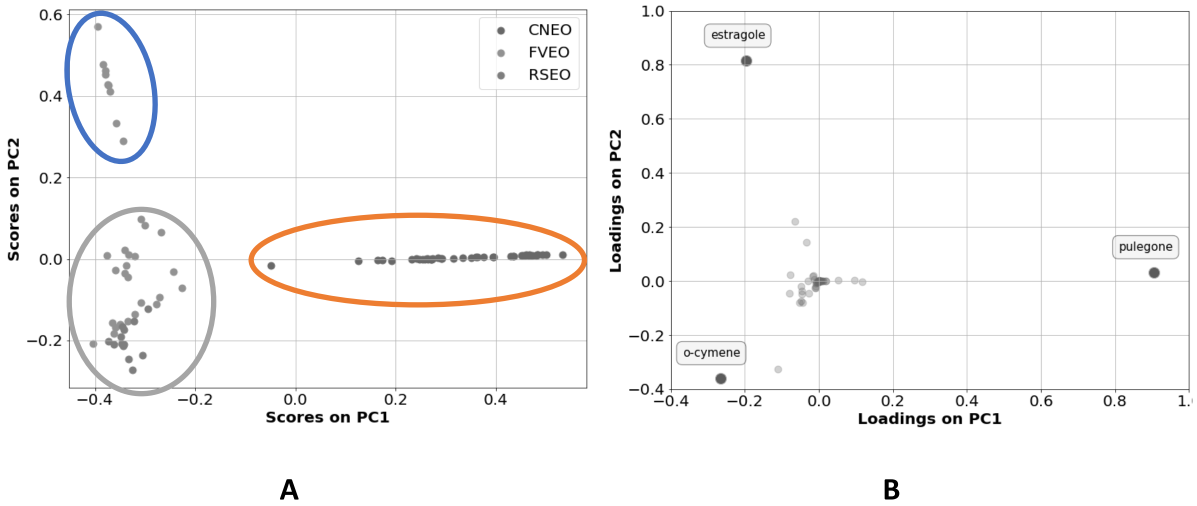


Figure 22. Graphical plot of the PCA's first two principal components. The scores plot (panel **A**) indicates at least three clusters (circled in panel **A**). The loadings plot (panel **B**) highlights that estragole, o-cymene, and pulegone could be the most important chemical constituents among all the tested EOs.

4.3.6 Binary Classification Models

4.3.6.1 General Results

Analogously as in ML application to *Pseudomonas aeruginosa* (PA), direct application of linear classification methods using algorithms such as Logistic Regression (LR) and Linear Support Vector Machines[7] (SVM) did not lead to satisfying classifiers (data not shown). At the same time, non-linear algorithms like random forest[8] (RF), non-linear support vector machine[102] (SVM), and gradient boosting[9] (GB) also led to insufficiently robust models (data not shown). Therefore, a mixed approach was used and taking the idea from the principal component regression (PCR) as an evolution of multiple linear regression (MLR) a number of PCs were used in place of the original variables (EOs chemical component percentages) as input for the sklearn LR implementation (PCLR). The PCLR was run on the PA dataset as an initial test, leading to highly overlapping results with those obtained with the GB application (data not shown). Nevertheless, as biofilm production assay profiled EOs as either inhibitors or activators (**Table 16**) accordingly, classification models were tentatively built for all four strains considering either biofilm production inhibition or activation for biofilm percentages observed at the two above introduced concentration levels of 48.8 µg/mL and 3.125 mg/mL. To this aim, initially, the optimal biofilm production percentage cutoff for the binary classification was explored by systematically either decreasing it from a starting 80% to 60% or increasing from 120% to 140% for the inhibition or activation models, respectively, being the ranges arbitrarily chosen on the basis of **Table 16** filled data. The models' accuracy was monitored by the MCC value obtained by leave-one-out cross-validation. Following this protocol, for the inhibition models, EOs samples characterized by higher values than the best performing cutoff of biofilm production percentage were classified as inactive, while those with lower values were considered active. On the contrary, regarding the biofilm production enhancer models, EOs samples characterized by higher percentages than the cutoff value were classified as active, while those with lower values were considered non-active.

Regarding the 6538P inhibition training set, the very low active/inactive ratio at biofilm inhibition below 80% prevented any optimization. Therefore, the grid search analysis to the starting sixteen training sets (four strains by two series of models by two concentrations) afforded to seven optimized models for either concentration (**Table 17**). Inspection of optimized models on both hyperparameters and cutoff values revealed for 25923/inhibition, RP62A/activation, and O-47/activation sets composed of high unbalanced ratios of actives over non-actives and were hence not further analyzed. Comparing developed models for the two used EOs concentrations revealed 3.125 mg/mL level to lead to more reliable and robust models (**Table 18**). Based on the above preliminary data, subsequent results and analyses were only carried out on RP62A/inhibition, O-47/inhibition, 6538P/activation, and 259237/activation models derived for biofilm modulation recorded at 3.125 mg/mL. This is in full agreement with the fact that EOs samples acted prevalently as a reducer of biofilm production for RP62A and O-47 strains, while for 6538P and 25923, the biofilm production was mainly enhanced (**Table 16**).

Table 17. Characteristics of the grid search optimized models.

Assayed Conc. (µg/mL)	Models' Parameters	Biofilm Inhibition Models			Biofilm Activation Models			
		RP62A	O-47	25923	RP62A	O-47	6538P	25923
3125	PCs [1]	5	19	22	9	12	9	25
	Actives [2]	31	30	4	6	4	27	20
	Non-actives [3]	58	59	85	83	85	62	69
	cutoff	62	62	63	126	133	139	139
48.8	PCs [1]	8	9	24	15	8	9	20
	Actives [2]	32	30	3	7	4	30	45
	Non-actives [3]	57	59	86	82	85	59	44
	Cutoff[4]	63	63	62	124	138	133	121

[1]: number of principal components used in the model;

[2]: number of EOs as inhibitors or enhancers of bacterial biofilm production;

[3]: number of EOs as non-inhibitors or not-enhancers of biofilm production;

[4]: optimal values of bacterial biofilm production percentage for binary classification as inhibitors/non-inhibitors or enhancers/not-enhancers of bacterial biofilm production;

Table 18. Fitted and cross-validated Accuracy, MCC, Precision-Recall and ROC-AUC coefficients for the RP62A/inhibition, O-47/inhibition, 6538P/activation and 259237/activation optimized models at 3.125 mg/mL and 0.0488 µg/mL.

Validation	Assayed Conc. (µg/mL)	Coefficient	Biofilm Inhibition Models		Biofilm Activation Models	
			RP62A	O-47	6538P	25923
Fitting	3125	Accuracy	0.721	0.771	0.832	0.906
		MCC	0.455	0.590	0.667	0.826
		Precision-Recall	0.657	0.682	0.772	0.956
		ROC-AUC	0.742	0.753	0.824	0.961
	48.8	Accuracy	0.722	0.780	0.806	0.763
		MCC	0.452	0.604	0.632	0.533
		Precision-Recall	0.659	0.681	0.757	0.824
		ROC-AUC	0.735	0.752	0.815	0.834
Cross Validation	3125	$\text{Accuracy}_{\text{CV}}$	0.687	0.738	0.805	0.784
		MCC_{CV}	0.392	0.517	0.613	0.568
		$\text{Precision-Recall}_{\text{CV}}$	0.584	0.589	0.698	0.782
		$\text{ROC-AUC}_{\text{CV}}$	0.683	0.659	0.743	0.845
	48.8	$\text{Accuracy}_{\text{CV}}$	0.663	0.721	0.722	0.606
		MCC_{CV}	0.335	0.474	0.450	0.214
		$\text{Precision-Recall}_{\text{CV}}$	0.577	0.591	0.668	0.533
		$\text{ROC-AUC}_{\text{CV}}$	0.666	0.660	0.753	0.599

Lack of chance correlation was checked to assess either models' fitness and robustness. Y-scrambling procedure whose 100 runs of the cross-validated scrambled set led to average, standard deviation, maximum and minimum values for $\text{Accuracy}_{\text{Y-S}}$, $\text{MCC}_{\text{Y-S}}$, $\text{Precision-Recall}_{\text{Y-S}}$, and $\text{ROC-AUC}_{\text{Y-S}}$ coefficients always lower than non-cross-validated and cross-validated ones, therefore assessing the validity of all final models (**Table 19**).

Table 19. Chance correlation control by Y-scrambling procedure results. Mean, standard deviation (St Dev), maximum (max) and minimum (min) values for Accuracy$_{Y-S}$, MCC$_{Y-S}$, Precision-Recall$_{Y-S,}$ and ROC-AUC$_{Y-S}$ ROC-AUC coefficients for cross-validated 100 runs. Values refer to RP62A/inhibition, O-47/inhibition, 6538P/activation and 259237/activation optimized models at 3.125 mg/mL.

Type of Model	Strain	Coefficient	Mean	St Dev	Max	Min
Biofilm Inhibition Models	RP62A	Accuracy$_{Y-S}$	0.500	0.079	0.644	0.219
		MCC$_{Y-S}$	0.000	0.159	0.290	−0.567
		Precision-Recall$_{Y-S}$	0.496	0.063	0.643	0.353
		ROC-AUC$_{Y-S}$	0.459	0.094	0.627	0.198
	O-47	Accuracy$_{Y-S}$	0.494	0.081	0.665	0.286
		MCC$_{Y-S}$	−0.011	0.164	0.342	−0.429
		Precision-Recall$_{Y-S}$	0.506	0.068	0.668	0.377
		ROC-AUC$_{Y-S}$	0.474	0.089	0.645	0.241
Biofilm Activation Models	6538P	Accuracy$_{Y-S}$	0.492	0.082	0.637	0.249
		MCC$_{Y-S}$	−0.017	0.169	0.275	−0.522
		Precision-Recall$_{Y-S}$	0.507	0.071	0.661	0.347
		ROC-AUC$_{Y-S}$	0.470	0.100	0.644	0.166
	25923	Accuracy$_{Y-S}$	0.508	0.083	0.680	0.292
		MCC$_{Y-S}$	0.013	0.170	0.361	−0.433
		Precision-Recall$_{Y-S}$	0.524	0.071	0.746	0.355
		ROC-AUC$_{Y-S}$	0.490	0.102	0.675	0.179

4.3.6.2 Binary Classification Model for 6538P Biofilm Production Activation

The 6538P/activation/3.125mg/mL optimized derived model was maximum at a cutoff of 133%, using 9 PCs, characterized by a 27:62 (0.44) proportion between actives and non-actives and high values of Accuracy (0.832), MCC (0.667), Precision-Recall (0.772) and ROC-AUC (0.824) coefficients which persisted to be quite good in cross-validation (Accuracy$_{CV}$ = 0.805, MCC$_{CV}$ = 0.613, Precision-Recall$_{CV}$ = 0.698 and ROC-AUC$_{CV}$ = 0.743) (**Table 18** and **Table 19**). Feature importance and partial dependence pointed out compounds 3-octanol, d-limonene and pulegone as more important for biofilm production enhancement modulation (**Figure 22**, **Figure 23**, **Figure 24** and **Table 20**).

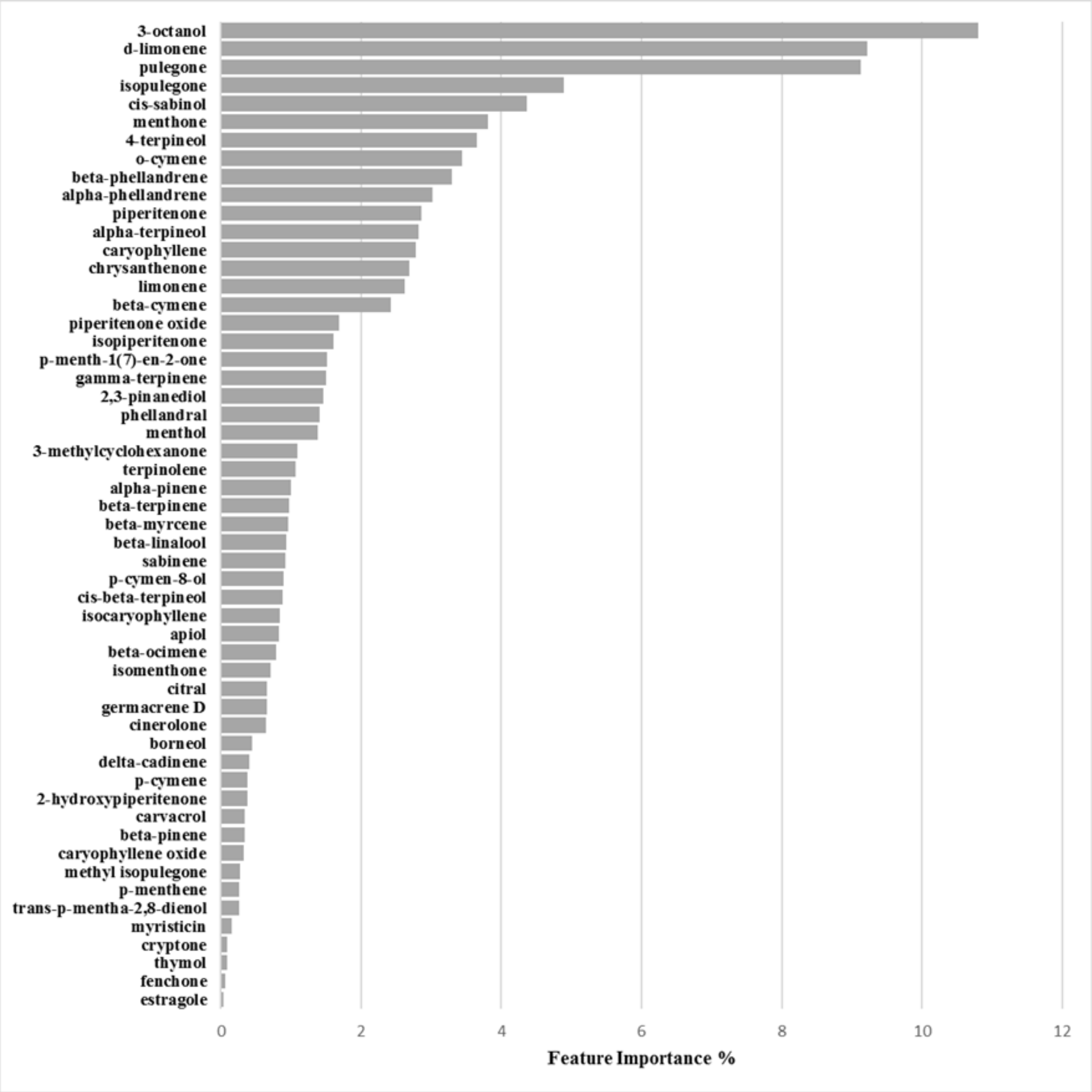


Figure 23. Feature importance plot for the 6538P/activation model defined at 3.125 mg/mL.

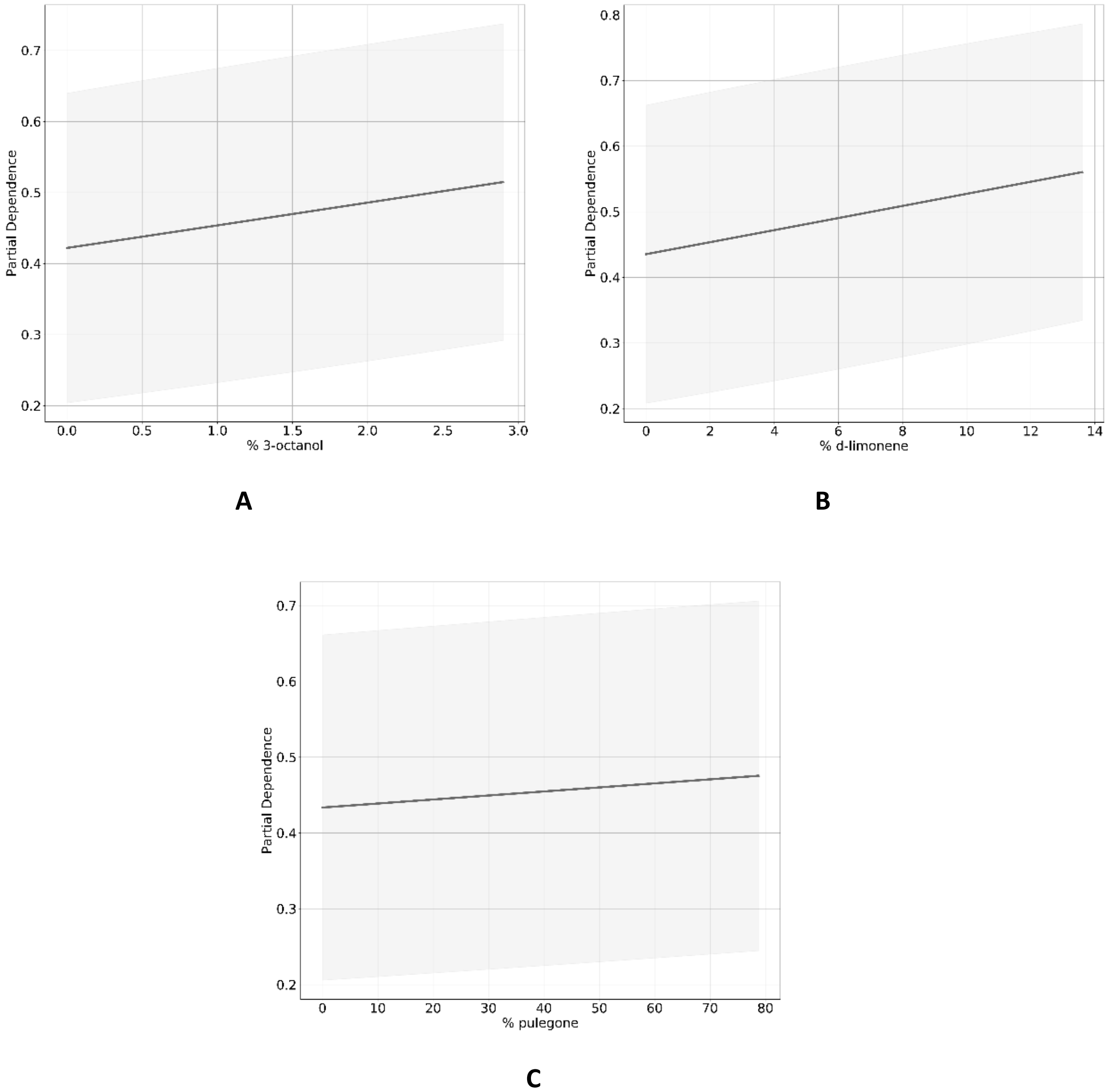


Figure 24. 3-octanol (**A**), d-limonene (**B**) and pulegone (**C**) partial dependence plots for the activation model on 6538P biofilm production.

4.3.6.3 Binary Classification Model for RP62A biofilm production inhibition

The grid search on the EOs' chemical composition and their associated RP62A biofilm production inhibitory potencies at 3.125 mg/mL identified 62% biofilm residual production as the best cutoff value with only 5 PCs and actives over non-actives ratio of 31:58 (0.53). The final classification model was found characterized by Accuracy, MCC, Precision-Recall, and ROC-AUC values of 0.721, 0.455, 0.657, and 0.742, respectively (**Table 18**). Cross-validation associated coefficients $\text{Accuracy}_{\text{CV}}$, MCC_{CV}, $\text{Precision-Recall}_{\text{CV}}$ and $\text{ROC-AUC}_{\text{CV}}$ were 0.687, 0.392, 0.584 and 0.683, respectively. Inspection of the model associated EOs' chemical components importance, the Skater algorithm indicated 3-octanol, phellandral, thymol, and d-limonene as those mostly influencing biofilm production inhibition (**Figure 25**, **Table 19**, and **Table 20**), whose positive control was highlighted through partial dependence plots which describe the marginal impact of a feature on model prediction (**Figure 26**).

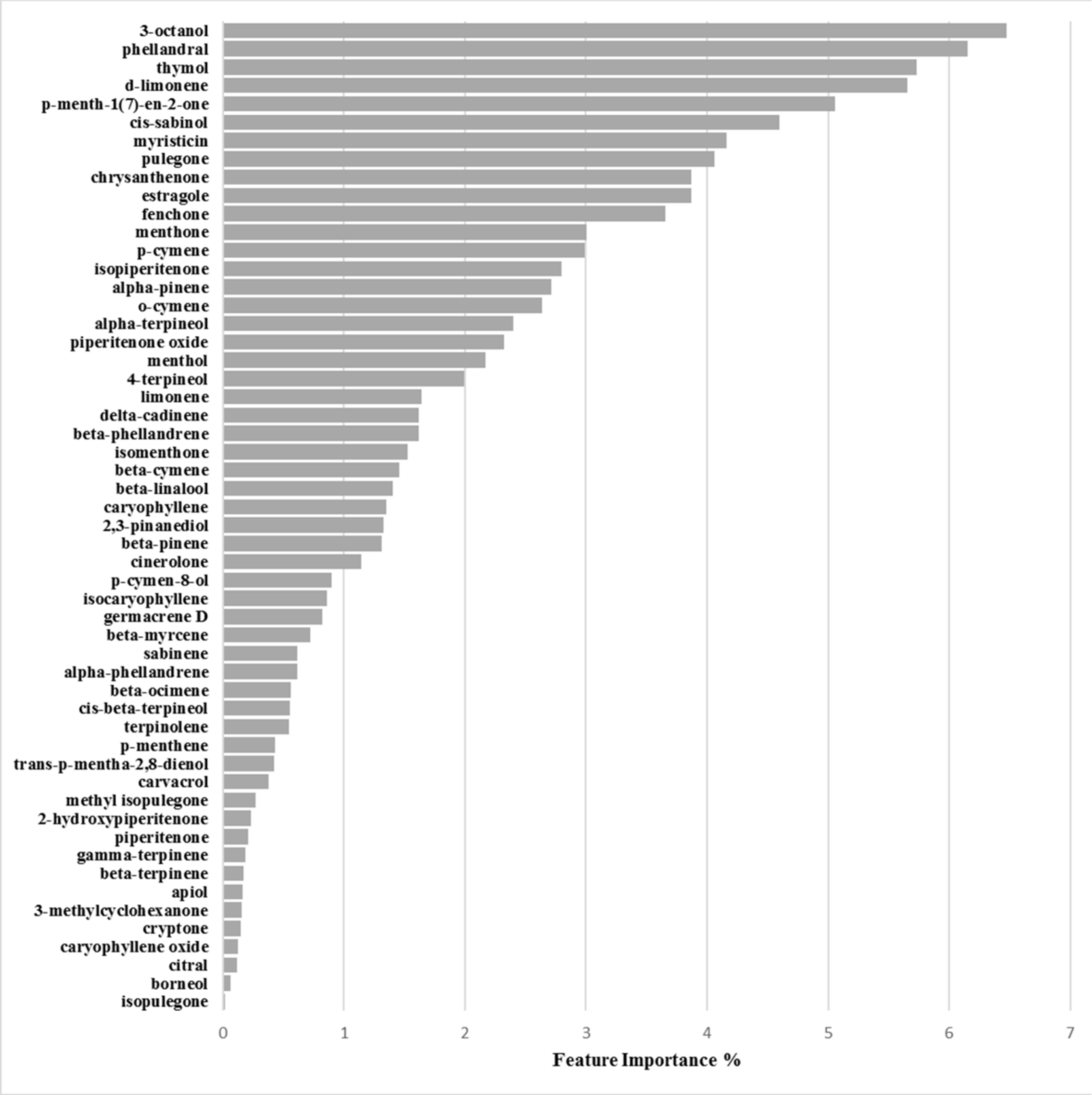


Figure 25. Feature importance plot for the RP62A/inhibition model defined at 3.125 mg/mL.

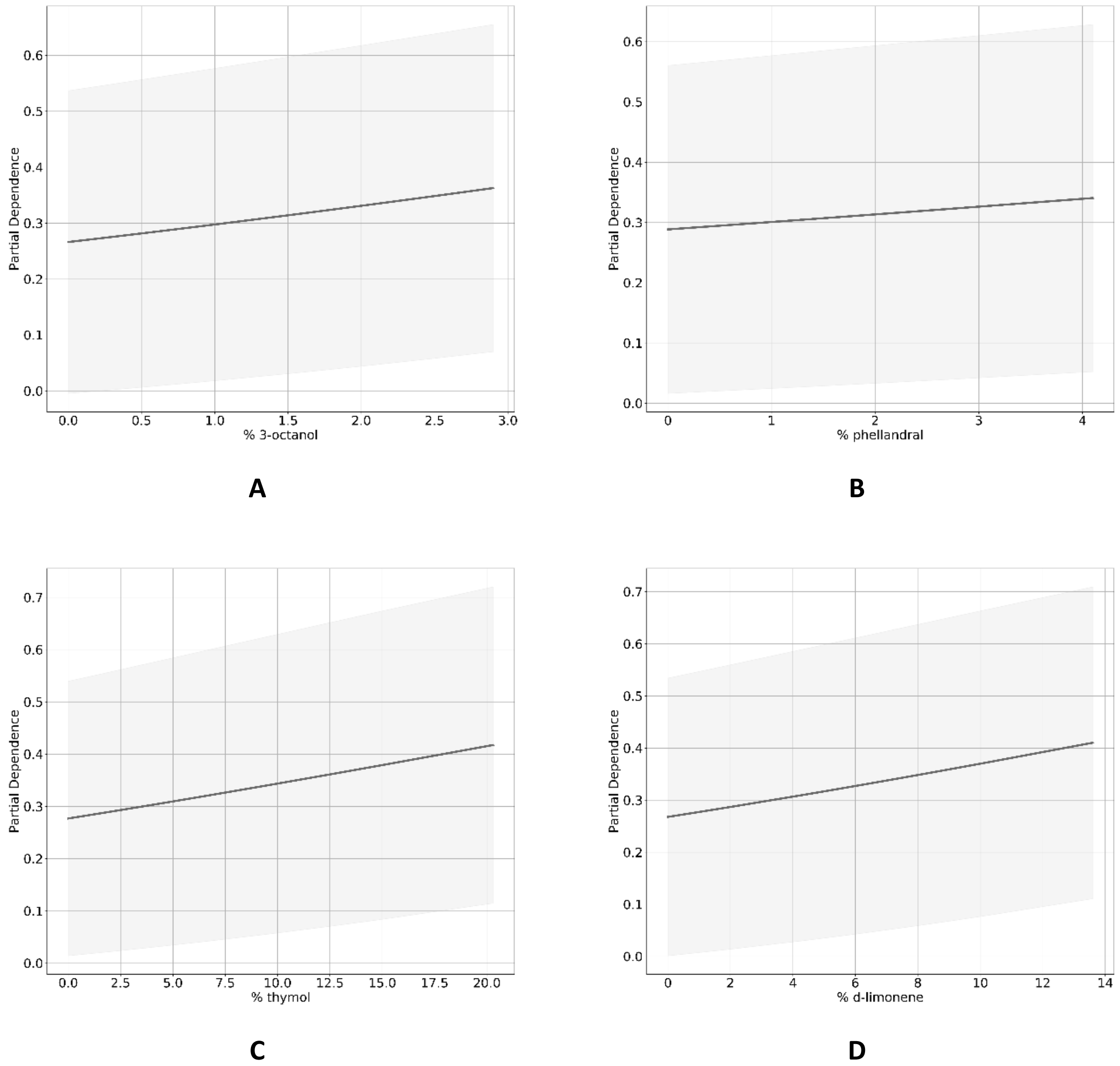


Figure 26. 3-octanol (**A**), phellandral (**B**), thymol (**C**), and d-limonene (**D**) partial dependence plots for the inhibition model on RP62A biofilm.

4.3.6.4 Binary Classification Model for O-47 Biofilm Production Inhibition

The O-47/inhibition/3.125mg/mL optimized derived model was also obtained at cutoff of 62%, but with 19 PCs and a 0.51 actives:non-actives ratio (30:59) leading to 0.771, 0.590, 0.682 and 0.753 values for the Accuracy, MCC, Precision-Recall and ROC-AUC coefficients, respectively. Model robustness was assessed by Accuracy$_{CV}$, MCC$_{CV}$, Precision-Recall$_{CV}$ and ROC-AUC$_{CV}$ values of 0.738, 0.517, 0.589 and 0.659, correspondingly (**Table 18**). The Y-scrambling application did not reveal the presence of any chance correlation (**Table 19**). Inspection of feature importance and partial dependence pointed out as more significant for biofilm production inhibition the compounds 3-octanol, o-cymene, d-limonene, and β-phellandrene (**Figure 28**, **Figure 28**, and **Table 20**).

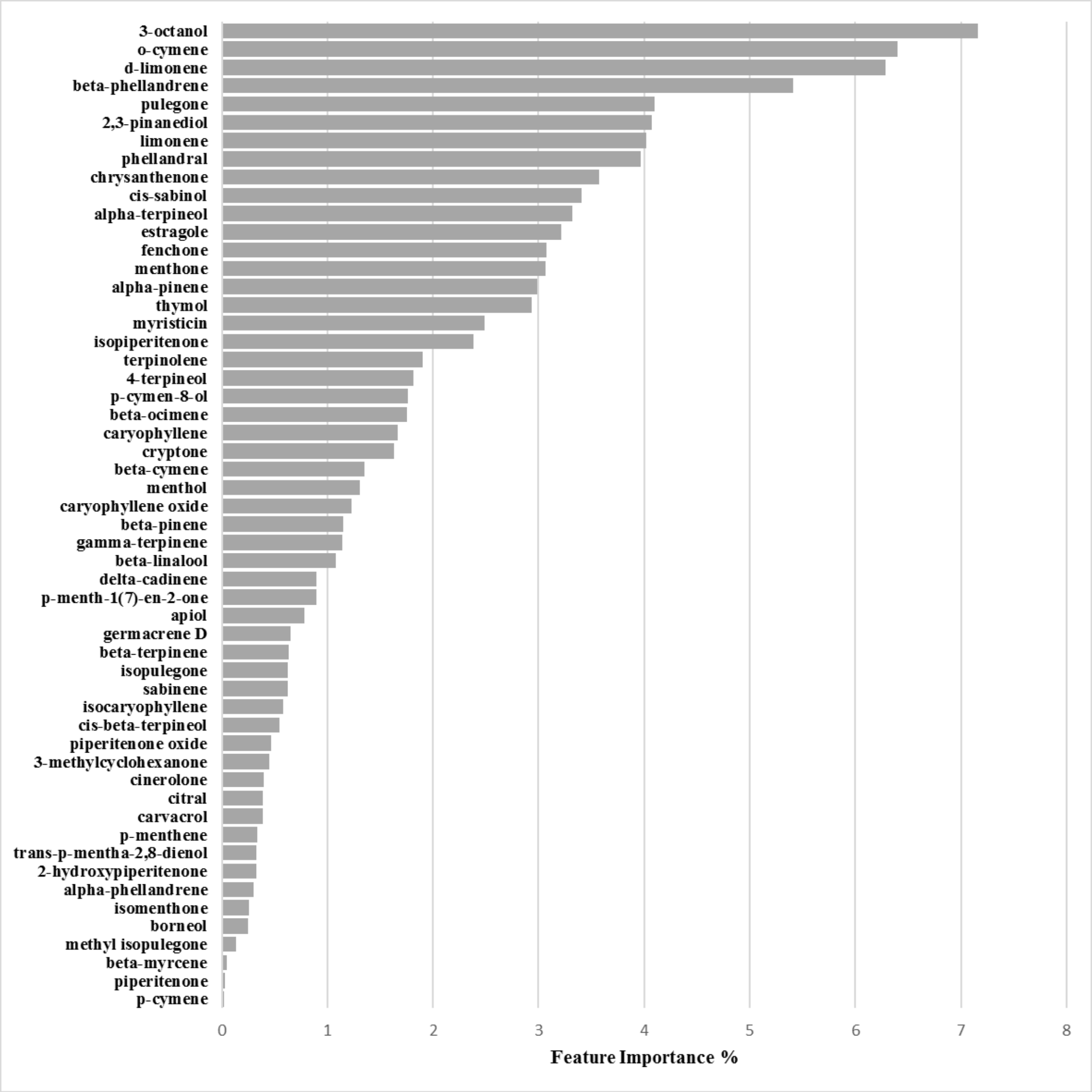


Figure 27. Feature importance plot for the O-47/inhibition model defined at 3.125 mg/mL.

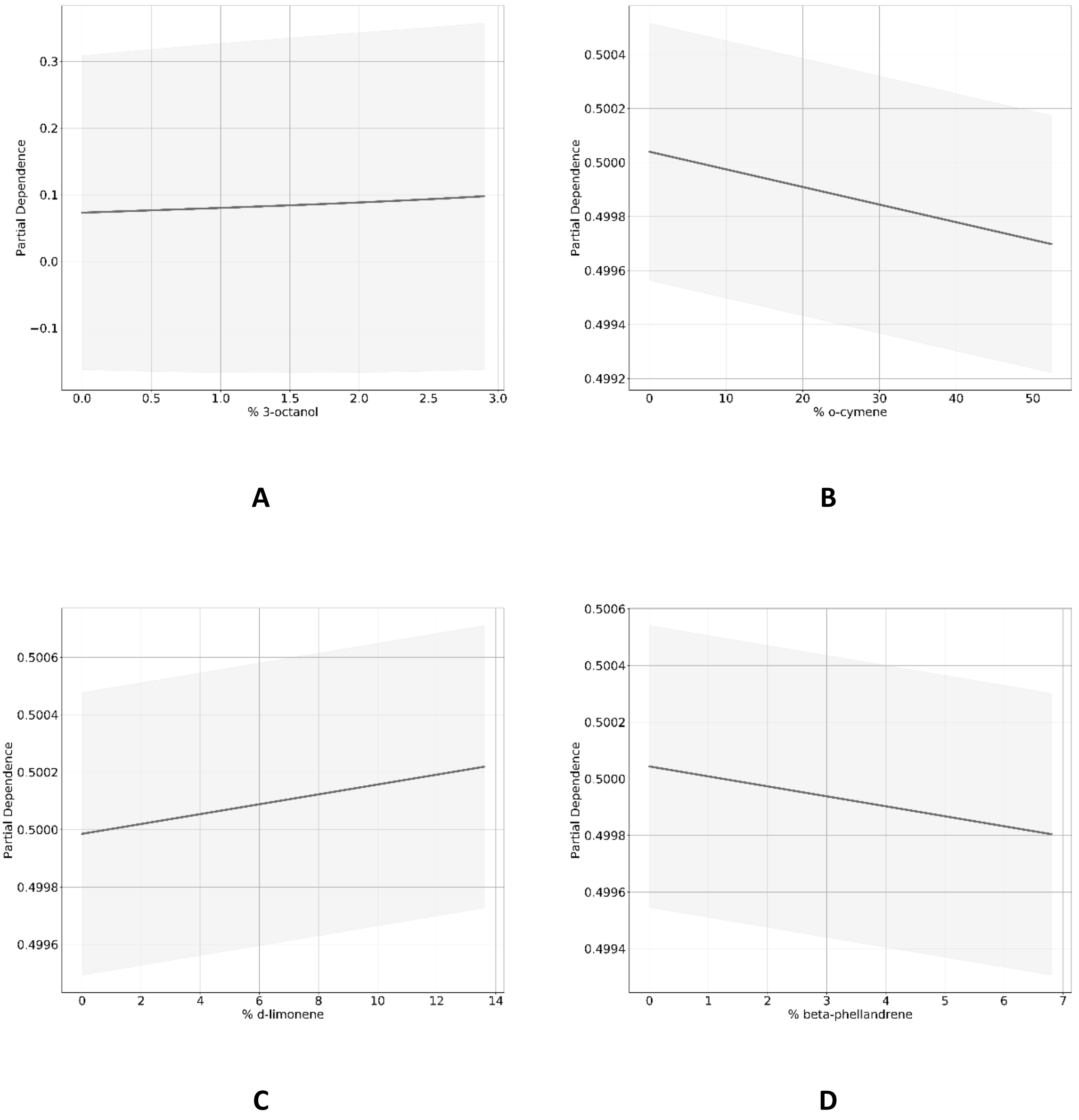


Figure 28. 3-octanol (**A**), o-cymene (**B**), d-limonene (**C**), and beta-phellandrene (**D**) partial dependence plots for the inhibition model on O-47 biofilm.

4.3.6.5 Binary Classification Model for 25923 Biofilm Production Activation

The 25923/activation/3.125mg/mL optimized derived model was determined at cutoff of 121%, using 25 PCs, characterized by a 20:69 (0.29) proportion between actives and non-actives. The non cross-validated model was characterized by Accuracy, MCC, Precision-Recall and ROC-AUC coefficients of 0.906, 0.826, 0.956 and 0.961, respectively. Model robustness by cross-validation was characterized by high values of Accuracy$_{CV}$ = 0.763, MCC$_{CV}$ = 0.533, Precision-Recall$_{CV}$ = 0.824 and ROC-AUC$_{CV}$ = 0.834 (**Table 18** and **Table 19**). Feature importance and partial dependence pointed out compounds menthone, menthol, β-linalool, β-cymene, chrysanthenone, 3-octanol as more important for biofilm production enhancement modulation (**Figure 29** and **Table 20**).

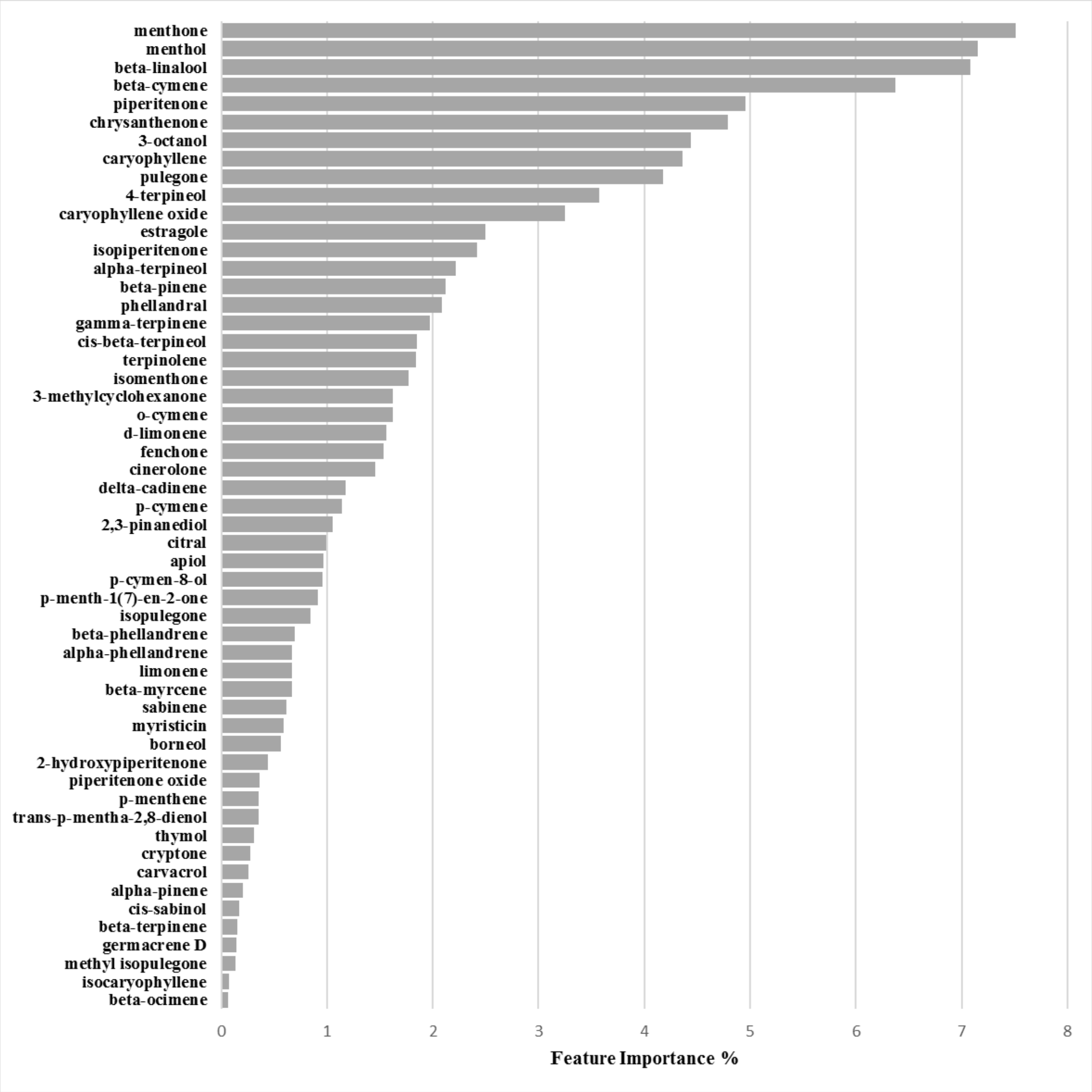


Figure 29. Feature importance plot for the 25923/activation model is defined at 3.125 mg/mL.

Table 20. Feature importances for each chemical component as derived by the SKATER algorithm for the RP62A/inhibition, O-47/inhibition, 6538P/activation, and 259237/activation optimized models at 3.125 mg/mL. Background of more important chemical components for inhibition are colored in darker green, while in the darker red background are highlighted components associated with higher biofilm production enhancement.

Chemical Component	**Biofilm Inhibition Models**		**Biofilm Inhibition Models**	
	RP62A	**O-47**	**6538P**	**25923**
2-Hydroxypiperitenone	0.23	0.32	0.38	0.44
2,3-Pinanediol	1.33	4.07	1.47	1.05
3-Methylcyclohexanone	0.16	0.45	1.09	1.62
3-Octanol	6.47	7.16	10.80	4.44
4-Terpineol	1.99	1.81	3.65	3.57
α-Phellandrene	0.61	0.30	3.02	0.67
α-Pinene	2.71	2.98	1.00	0.21
α-Terpineol	2.40	3.32	2.81	2.22
Apiol	0.17	0.78	0.83	0.97
β-Cymene	1.46	1.35	2.42	6.37
β-Linalool	1.40	1.08	0.93	7.08
β-Myrcene	0.72	0.05	0.95	0.67
β-Ocimene	0.56	1.75	0.78	0.06
β-Phellandrene	1.62	5.42	3.30	0.69
β-Pinene	1.31	1.15	0.34	2.12
β-Terpinene	0.17	0.63	0.97	0.15
Borneol	0.07	0.25	0.45	0.56
Carvacrol	0.38	0.38	0.34	0.26
Caryophyllene	1.35	1.67	2.78	4.36
Caryophyllene oxide	0.13	1.23	0.32	3.25
Chrysanthenone	3.87	3.58	2.68	4.79

Table 20. Feature importances for each chemical component as derived by the SKATER algorithm for the RP62A/inhibition, O-47/inhibition, 6538P/activation, and 259237/activation optimized models at 3.125 mg/mL. Background of more important chemical components for inhibition are colored in darker green, while in the darker red background are highlighted components associated with higher biofilm production enhancement.

Chemical Component	**Biofilm Inhibition Models**		**Biofilm Inhibition Models**	
	RP62A	**O-47**	**6538P**	**25923**
Cinerolone	1.14	0.40	0.64	1.45
cis-β-Terpineol	0.55	0.55	0.88	1.85
cis-Sabinol	4.60	3.40	4.36	0.17
Citral	0.12	0.39	0.65	0.99
Cryptone	0.15	1.63	0.09	0.27
d-Limonene	5.66	6.29	9.22	1.56
delta-Cadinene	1.62	0.90	0.40	1.17
Estragole	3.87	3.21	0.04	2.49
Fenchone	3.66	3.08	0.06	1.54
γ-Terpinene	0.19	1.14	1.51	1.97
Germacrene D	0.82	0.65	0.65	0.14
Isocaryophyllene	0.86	0.58	0.84	0.07
Isomenthone	1.53	0.26	0.71	1.77
Isopiperitenone	2.80	2.38	1.61	2.42
Isopulegone	0.01	0.63	4.89	0.85
Limonene	1.64	4.02	2.62	0.67
Menthol	2.17	1.30	1.38	7.16
Menthone	3.00	3.07	3.81	7.51
Methyl isopulegone	0.27	0.14	0.27	0.13
Myristicin	4.16	2.48	0.16	0.59
o-Cymene	2.64	6.40	3.44	1.62

Table 20. Feature importances for each chemical component as derived by the SKATER algorithm for the RP62A/inhibition, O-47/inhibition, 6538P/activation, and 259237/activation optimized models at 3.125 mg/mL. Background of more important chemical components for inhibition are colored in darker green, while in the darker red background are highlighted components associated with higher biofilm production enhancement.

Chemical Component	Biofilm Inhibition Models		Biofilm Inhibition Models	
	RP62A	O-47	6538P	25923
p-Cymen-8-ol	0.90	1.76	0.89	0.96
p-Cymene	2.99	0.02	0.39	1.14
p-Menth-1(7)-en-2-one	5.06	0.89	1.51	0.91
p-Menthene	0.43	0.33	0.26	0.35
Phellandral	6.15	3.97	1.40	2.08
Piperitenone	0.21	0.03	2.86	4.95
Piperitenone oxide	2.32	0.46	1.68	0.36
Pulegone	4.06	4.10	9.12	4.18
Sabinene	0.61	0.62	0.91	0.62
Terpinolene	0.55	1.90	1.07	1.84
Thymol	5.73	2.94	0.08	0.31
trans-*p*-Mentha-2,8-dienol	0.43	0.33	0.26	0.35

4.4 Discussion and Conclusions

4.4.1 EOs Biofilm Bioactivity General Consideration

From the results reported above, it could be observed that each EO had a specific effect on biofilm formation, likely depending on its characteristics and unique chemical composition. In particular, for *S. aureus* strains 6538P and 25923, the EOs mainly exhibited an enhancement of biofilm production. Stimulation of bacterial biofilm production by EOs is not surprising as it was previously observed, even by isolated chemical components[133–135]. On the other hand, and more common[136,137], for *S. epidermidis* strains RP62A and O-47, an overall inhibition effect on biofilm production was observed by in vitro EOs treatment. Nevertheless, cinnamon

EO was reported to stimulate biofilm production on some *Staphylococcus epidermidis* strains[138].

4.4.2 Bioactivity of RSEOs

The majority of tested RSEO samples did not show inhibitory effects on *S. aureus* 6538P biofilm formation (a partial inhibitory effect was observed only for R6 essential oil at 3.125 mg/mL, panel A of **Figure 16**). On the contrary, some RSEO samples (R6, R12, R24, RM4, and RM6) were shown to enhance biofilm formation by up to 140% at 0.0488 mg/mL. Differently, several RSEO samples showed a good inhibitory effect on *S. epidermids* RP62A biofilm production. In particular, 4 out of 13 EOs (R6, R24, RM4, and RM6) were able to potently inhibit biofilm formation with a rate of about 80% at either used concentrations (panel B of **Figure 16**). Thus, these EOs were selected for further analyses using each EO's scalar concentration, starting from 0.0488 mg/mL. An attempt to determine a direct dose-dependent effect was not effective (**Figure 19**).

On O-47 biofilm modulation (panel C of **Figure 16**), most RSEOs had a slight inhibition effect of up to 40% (60% residual of biofilm production) at 0.0488 mg/mL. On the contrary, at the higher concentration, RSEOs enhanced biofilm production by up to 130% for most samples. Only RM6 showed a remarkable biofilm production up to 160% at 3.125 mg/mL. For strain 25923, a profile similar to that of RP62A was observed (panel D of **Figure 16**). Therefore, no further investigation was pursued on the R6, R24, RM4, and RM6 samples despite the high inhibitory biofilm potency with residual biofilm production ranging 20-30%.

4.4.3 Bioactivity of FVEOs

Among all tested EOs, those from FV comprise among the most active samples able to inhibit biofilm production (**Figure 17**). In particular, only mild effects (positive or negative modulation of biofilm production) were observed on 6538P strain (panel A of **Figure 17**) with a few exceptions at both selected concentrations, including FA24, FS2, FO3, and FO6 that increased biofilm production by about 40–60%, and FOM1 that inhibited biofilm production by about 50% at 3.125 mg/mL. On the contrary, some FVEOs proved to be potent antibiofilm agents on *S. epidermids* RP62A (panel B of **Figure 17**). In particular, 5 out 33 FVEOs (FO1, FO3, FO6, FO24, and FOM3) inhibited biofilm formation with a rate of about 80% (panel B of **Figure 17**) and were selected for further analyses using a scalar concentration of each EO starting from 0.0488 mg/mL. Similarly, as for the selected potent RSEOs, no direct dose-dependent effect was determined (**Figure 20**). Interestingly, on O-47 biofilm modulation, most FVEOs showed a bioactivity profile almost overlapping that for RP62A with FO1, FO3, FO6, FO24, and FOM3 samples able to reduce biofilm formation of about 50-70% (panel C of **Figure 17**). While on both *S. epidermidis* strains RP62A and O-47, FVEOs displayed some peculiar samples with interesting biofilm inhibitory potencies in the case of 25923 strain FVEOs displayed an overall bioactivity profile similar to that for RSEOs against O-47 (compare panel C of **Figure 16** with panel D of **Figure 17**).

4.4.4 Bioactivity of CGEOs

EOs from CG are the most modulating biofilm producers either in positive (activators) or in negative (inhibitors). In particular, in the case of strain 6538P, all samples at either concentration can be classified as neutral or biofilm promoters (panel A of **Figure 18**) with a strong inclination to increase biofilm production by up to 500% (CAM1). Other strong biofilm inducers (percentages over 300%) are CAM3, CAM5, CS1, CS3, CS6, and CS24 samples. Many other CGEOs, although to a lesser extent, induced a doubling or even tripling of biofilm production. On the contrary, most of CGEOs displayed an inhibition by over 50% of biofilm production by RP62A (panel B of **Figure 18**). Many CGEO samples were further investigated, and for 6 of them (CO2, CO6, COM5, CS2, CS6, and CSM5), a definite dose-dependent relation was observed (**Figure 21**). Regarding biofilm modulation for O-47 CGEOs, in this case, presented, at either tested concentrations, a mixed scenario in which some samples induced an enhanced biofilm production up to 250-350% (CAM5, CS1, CS3, CS6, CSM1, and CSM3) and 15 different samples showed high inhibition potencies (percentages of residual biofilm lower than 40–50%).

4.4.5 Machine Learning Classification Models

Application of the PCA coupled with logistic regression led to the formulation of 4 robust models that were characterized by good Accuracy, MCC, Precision-Recall, and ROC-AUC values (**Table 19**). Model agnostic feature importance and partial dependence plots were used to find the marginal effect that each EO chemical component has on the predicted outcome of the binary classification models built on the 3.125 mg/mL response variables. Feature importance is a measure of the prediction error of the model after the feature's values are permuted and highlights the absolute importance of each chemical constituent, while partial dependence plots show whether the relationship between the bioactivity and the chemical component is linear, monotonous, or more complex.

4.4.6 Biofilm Activation ML Model on 6538P

Inspection of feature importance for model derived on 6538P biofilm percentage production, and EOs' chemical compositions revealed 3-octanol, d-limonene, and pulegone as the chemical components more associated with bacterial biofilm production (**Figure 23** and **Table 20**). Further investigation of their partial dependence plots (**Figure 24**) indicated those three chemicals as all positively correlated with biofilm enhancement.

4.4.7 Biofilm Activation ML Model on 25923

Similarly, as for 6538P, also for the 25923 strain, an ML model was built to correlate biofilm production enhancement with EOs' chemical composition. Again, analysis of feature importance was found as more important menthone, menthol, β-linalool, β-cymene, chrysanthenone, 3-octanol (**Figure 29**). Differently, as found for 6538P, the main component was not all positively associated with biofilm enhancement production. Feature importance and partial dependence pointed out compounds menthone, menthol, β-linalool, β-cymene, chrysanthenone, 3-octanol as more critical for biofilm production enhancement modulation (**Figure 29** and **Table 20**).

4.4.8 Biofilm Inhibition ML Model on RP62A

Unlike the previous model, feature importance associated with the EOs biofilm inhibition production on RP62A strain highlighted 3-octanol, phellandral, thymol, and d-limonene as chemical compounds important on modulating biofilm reduction (**Figure 25**, **Table 19**, and **Table 20**). Partial dependence plots for 3-octanol, phellandral, thymol, d-limonene associated the four components as all positively able to inhibit biofilm production (**Figure 26**).

4.4.9 Biofilm Inhibition ML Model on O-47

Regarding the ML model derived on the biofilm inhibition capability of EOs, the compounds more responsible for biofilm production modulation were found to be 3-octanol, o-cymene, d-limonene, and β-phellandrene (**Figure 28** and **Table 20**). Differently from the above RP62A analogous inhibition model, only 3-octanol and d-limonene were found positively associated with EOs' inhibitory ability by partial dependence plots (**Figure 28**). On the contrary, o-cymene and β-phellandrene were associated with negative action on the inhibition. This could be speculated as a sort of anti-synergic effect that could balance EOs' potencies.

4.4.10 General Consideration on ML Models

According to the four classification models, two compounds, namely 3-octanol and d-limonene, can be considered as those that most influence biofilm production (**Table 20**). In particular, d-limonene positively correlated either in inhibiting or enhancing biofilm production in three out of the four models while has a negative modulation on the ML model built on the biofilm enhancement of EOs' on 25923 strain. These data indicate some controversial mechanisms associated with d-limonene. It could be speculated that being this compound a highly apolar monoterpene, its role could not be indirectly associated to biofilm modulation by altering the bacterial wall[139] allowing other compounds, likely oxygenated ones, to enter the cell acting in altering some biochemical mechanism that could end in stimulation or inhibition of biofilm production. Nevertheless, on this topic, the data available in the literature is controversial: Natcha and Caoili[140] reported that d-limonene is effective in inhibiting the growth of *S. epidermidis* RP62A when combined with the antibiotic rifampicin, likely due to d-limonene interference with biofilm formation. The effect of d-limonene in inhibiting bacterial biofilm formation was also proved against species of the genus Streptococcus[141] for which minimal biofilm inhibitory concentration (MBIC) of 400 µg/mL was determined. In a very recent study, d-limonene was also reported as a biofilm inhibitor, although less efficient than an EO containing d-limonene[142]. On the contrary, Kerekes et al. assayed a series of EOs and a list of chemical components against food-related micro-organisms and found d-limonene was almost deprived of any ability to inhibit biofilm production. In a study from Espina et al., d-limonene at 2000 µL/L was reported to reduce biofilm mass production in *S. aureus* USA300 by 90% after 8 hours of incubation, but increase it by 30% after 40 h of incubation[143]. EOs containing d-limonene and the isolated component were found to stimulate biofilm production on *Listeria monocytogenes* and antibiotic-resistant *Enterococcus faecalis* strains[133–135]. A similar profile and speculation on 3-octanol could also be deduced. 3-Octanol is a molecule resembling normal octanol, a compound commonly used to evaluate compound membrane permeability and lipophilicity by determining the logP parameter often used in ADME and QSAR studies. No data are available on the influence of 3-octanol on biofilm production, except for a single report in which the 8-carbon molecules 1-octen-3-ol, 3-octanol and 3-octanone specifically induced conidiation in Trichoderma species colonies placed in the dark[144]. Pulegone, γ-terpinene, and piperitenone, could be the main components responsible for the modulation of EOs' augmented biofilm production for strains 6538P and 25923, based on the ML elaboration and considering the possible cell wall permeation role of both d-limonene and 3-octanol on these strains. In the case RP62A and O-47 phellandral, thymol, o-cymene, and β-phellandrene are mainly responsible for positively (phellandral and thymol) or negatively (o-cymene and β-phellandrene) modulating EOs' biofilm inhibition. Unfortunately, no specific data are available on these isolated components, and the herein discussion, although based on robust ML calculation are not experimentally based. It is worthy to note that the four bacterial strains tested here produced biofilms with different characteristics. First 6538P and 25923 belong to *S. aureus* species, while RP62A and O47 belong to *S. epidermidis* species. 25923 is classified as a strong biofilm producer, and 6538P is a medium/strong biofilm producer according to Cafiso and coworkers[145]. Proteins are the major component in the biofilm matrix of 6538P, while in 25923, the polysaccharides have a predominant role. As regards the *S. epidermidis* strains, they are both strong biofilm makers and produce a biofilm mainly composed of polysaccharides. Moreover, O-47 is a naturally occurring agr mutant[126]. As previously reported[146], agr-negative genotype enhanced biofilm formation on polymer surfaces by an increased expression of the surface protein AtlE, a bifunctional adhesin/autolysin abundant in the cell wall of *S. epidermidis*. The amount of AtlE present in cell envelop one of the reported differences between RP62A and O-47[146]. The overexpression of AtlE could induce significant changes in the hydrophobicity of the bacterial surface[147]; this effect could explain the different actions of EOs on these two strains.

Furthermore, the classification models were developed on the same EOs tested on *P. aeruginosa* biofilm production. In that case, investigation of the most important components through feature importance and partial dependence plots indicated estragole and phellandral as the chemical components mostly related to biofilm inhibition of *P. aeruginosa*. Concurrently, d-limonene, pulegone, and chrysanthenone seem to be related to its biofilm production. Although the use of feature importance and partial dependence plots shed some light on some EOs’ components' possible role, little is yet known on the role of the whole EOs mixture synergisms and anti-synergisms. Further studies on isolated EOs’ chemical components and their simple mixture are currently under evaluation to develop more refined ML models able to disclose more details on the EOs’ mechanism of action.

5 Essential oils against bacterial isolates from cystic fibrosis patients using antimicrobial and unsupervised machine learning approaches

5.1 Introduction

Cystic fibrosis (CF) is one of the most common lethal genetic disorders in the Caucasian population. It is inherited as an autosomal recessive disease and affects 70.000 persons worldwide (Cystic Fibrosis Foundation, CFF). The defective gene, identified in 1989, is the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) that is carried by 4% of persons (among Caucasians). Since CFTR encodes for a chloride channel of the epithelial cell surface, CF patients manifest a variety of multi-organ problems due to the alteration of sodium and chloride secretion across cell membranes and the subsequent luminal dehydration[148]. The impairment of mucociliary clearance, which should remove all microbes entering the airways, leads to the production of thick and dehydrated mucus in the CF lung, which promotes the airway chronic bacterial colonization[149].

The microbiology of the CF respiratory tract is peculiar. In the early stage of life, it is characterized by the prevalence of the Gram-positive bacterium *Staphylococcus aureus* (*S. aureus*). Overall, in 2017 more than half of the affected individuals had at least one culture positive for methicillin-sensitive *S. aureus* (MSSA). The highest prevalence of methicillin-resistant *S. aureus* (MRSA) occurs in individuals between the ages of 10 and 30, while MSSA reaches the peak among patients younger than 10 (Cystic Fibrosis Foundation. 2017. Patient Registry Annual Data Report https://www.cff.org/Research/Researcher-Resources/Patient-Registry/2017-Patient-Registry-Annual-Data-Report.pdf).

In early adolescence, CF patients' lung becomes chronically infected with Gram-negative non-fermenting bacteria. Among these, Pseudomonas aeruginosa (*P. aeruginosa*) is the most relevant and recurring, so that 30% of CF children and up to 80% of CF adults (25 years old and older) have lungs chronically colonized by this pathogen[150]. *P. aeruginosa* isolated from respiratory secretions demonstrates great phenotypic diversity and develops genetic mutations over time to adapt and survive in the complex environment of the CF airway[151]. *P. aeruginosa* mucoid phenotype, defined by the exopolysaccharide alginate overproduction within the lungs of CF patients, is a hallmark of chronic infection and predictive of poor prognosis. Indeed, mucoid *P. aeruginosa* has also been associated with failure of eradication and, compared to non-mucoid counterpart, exhibits enhanced resistance to multiple antibiotics and host immune effectors[152].

CF patients' life expectancy has consistently grown, reaching a median life of 40 years due to current treatments. CF patients born in 2010 are expected to live up to 50 years of age[153], assuming a positive trend of clinical care improvements at the actual rate.

The intensive use of antimicrobial drugs to fight lung infections inevitably leads to the onset of antibiotic-resistant bacterial strains. New antimicrobial compounds should be identified to overcome antibiotic resistance during the treatment of CF lung infections.

Recent investigation has disclosed a few small molecules, such as peptides or mannosides, showing promising efficacy in preventing and treating both bacterial and fungal biofilm infections *in vivo*[108]. Nevertheless, small molecules are known to select more and more resistant strains due to their mechanism of action based on specific binding to a primary target[154]. Interestingly, in recent literature, some reports on the use of naturally derived compounds showed in vitro the potentiality to inhibit the development of CF associated infections[43,44,93,155]. In particular essential oils seemed to be the most promising agents among tested natural compounds[43,44]. This study reports an extensive study on 61 essential oils (EOs) against a panel of 40 bacterial strains isolated from CF patients (see **Table 21**).

Table 21. Classification of bacterial strains based on their biofilm formation ability. For S. aureus, results were analyzed according to Cafiso et al.[145]; for *P. aeruginosa*, classification was based on Perez et al.[156]. NP: non biofilm producer. *Reference strains.

Bacterial strains	Biofilm producer	Bacterial strains	Biofilm producer
6538P*	STRONG	PAO1*	STRONG
25923*	STRONG	PA14*	STRONG
1S	WEAK	21P	STRONG
2S	MODERATE	22P	NP
3S	WEAK	23P	MODERATE
4S	WEAK	24P	NP
5S	WEAK	25P	WEAK
6S	WEAK	26P	NP
7S	MODERATE	27P	NP
8S	WEAK	28P	NP
9S	WEAK	29P	NP
10S	MODERATE	30P	WEAK
11S	WEAK	31P	MODERATE
12S	WEAK	32P	WEAK
13S	WEAK	33P	NP
14S	WEAK	34P	WEAK
15S	MODERATE	35P	WEAK
16S	WEAK	36P	WEAK
17S	MODERATE	37P	STRONG
18S	WEAK	38P	MODERATE
19S	WEAK	39P	WEAK
20S	MODERATE	40P	WEAK

The workflow in **Figure *30*** was followed to reduce in vitro procedure and render the investigation as convergent as possible. Unsupervised machine learning algorithms and techniques, as implemented in python language[128], were first applied to pick-up fewer representative strains (RS) among the panel of 40. To this aim, several categorical descriptors were collected and used to cluster the CF isolated strains. The clusters' centroids indicated the RS to be investigated for their susceptibility to a list of commercial EOs at fixed doses. Three EOs showed a great efficacy to reduce the microorganisms' growth and were therefore promptly assayed against all the available clinical isolates. The three EOs confirmed the initial assumption demonstrating their ability to inhibit bacterial growth. Gas chromatography, coupled with mass spectrometry (GC/MS), was then performed on the three EOs to investigate the likely chemical components mainly responsible for the antibacterial activity.

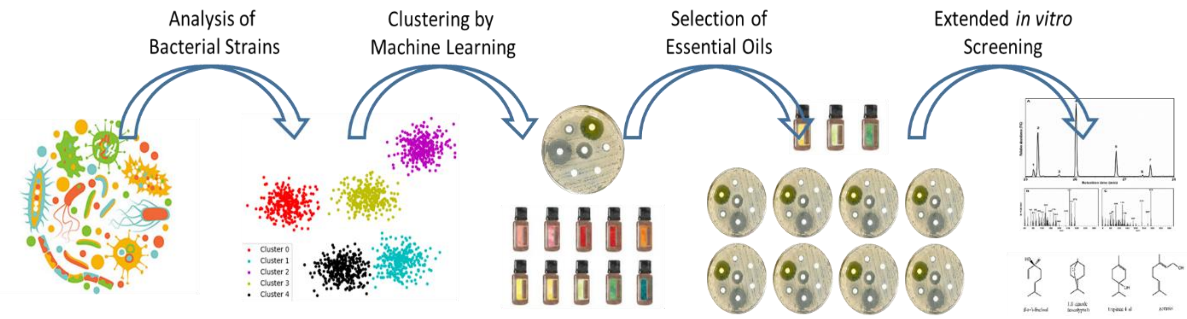

Analysis of Bacterial Strains
Clustering by Machine Learning
Selection of Essential Oils
Extended *in vitro* Screening
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4

Figure 30. The workflow of the herein investigation.

5.2 Material and Methods

5.2.1 Ethics approval and informed consent

The approval for this research was granted by the Ethics Committee of Children's Hospital and Institute Research Bambino Gesù in Rome, Italy (No 1437_OPBG_2017 of July 2017), and it was performed according to the principles of the Helsinki Declaration. Informed consent was obtained from all individual participants and all parents/legal guardians included in the study.

5.2.2 Clinical isolates from CF patients

In this study were used 40 bacterial strains (20 *S. aureus*, 20 *P. aeruginosa*) obtained from respiratory specimens of 30 CF patients (13 males, 17 females; average age 20.5) in follow-up at Pediatric Hospital Bambino Gesù (OPBG) of Rome, Italy. In particular, 27 bacterial strains were isolated from sputum, 11 from hypopharyngeal suction, and two from throat swabs (**Table 22** and **Table 23**). As reference strains were used: *S. aureus* ATCC 6538P (6538P) and *S. aureus* ATCC 25923 (25923), commonly recognized as a reference strain for antimicrobial testing; *P. aeruginosa* PAO1 (PAO1) and *P. aeruginosa* PA14 respectively recognized as moderately and highly virulent[157].

Table 22. The 20 Staphylococcus aureus clinical isolates and their characterization by several properties.

ID pt	ID	SAM	Date	Str	Ph	QUIN	B	ER	CLI	LIN	RCLI	CF	CPA	GEN
1	1S	ESP	10/11/2006	MRSA	SCV	R	S	Nt	R	S	N	Cp		J
2	2S	ESP	11/22/2007	MRSA	SCV	R	S	Nt	R	S	N	Ca	X	N
3	3S	ESP	1/15/2009	MRSA	SCV	S	S	Nt	S	S	N		X	E
4	4S	AT	2/20/2009	MRSA	—	S	S	Nt	S	S	P			A
5	5S	ESP	11/13/2009	MRSA	—	R	S	Nt	R	S	N	Sp		C
6	6S	AT	1/10/2011	MRSA	—	R	S	Nt	R	S	P			K
7	7S	ESP	4/4/2011	MRSA	—	R	S	Nt	R	S	N	Ca	X	D
8	8S	AT	7/22/2013	MRSA	—	R	S	Nt	S	S	N			I
9	9S	ESP	1/15/2014	MRSA	—	S	S	Nt	R	S	P	Ca	X	C
10	10S	AT	1/29/2015	MRSA	—	S	S	Nt	R	S	N	Ca/Cd/Pb		G
11	11S	AT	6/15/2017	MSSA	—	S	S	R	R	S	P			C
12	12S	AT	6/15/2017	MSSA	—	S	S	R	R	S	P			U
13	13S	AT	5/23/2017	MSSA	—	I	S	I	I	S	N	Sa		B
14	14S	AT	5/25/2017	MSSA	—	S	S	S	S	S	N			C
15	15S	AT	5/24/2017	MSSA	—	S	S	R	S	S	N		X	C
16	16S	AT	5/26/2017	MSSA	—	S	S	R	R	S	N			H
17	17S	AT	5/25/2017	MSSA	—	S	S	R	R	S	N	Af		M
18	18S	ESP	5/24/2017	MSSA	—	S	S	R	R	S	P	Ca	X	C
19	19S	ESP	6/15/2017	MSSA	—	S	S	R	R	S	P		X	L
20	20S	ESP	5/19/2017	MSSA	—	R	S	R	R	S	P		X	F

ID pt: patient identification; ID: strain code; SAM:Sample; Date: Date of the collection; Str:Strain; Ph: phenotype; QUIN: quinolones; B: Trimethoprim/Sulfamethoxazole; ER: Erythromycin; CLI: Clindamycin; LIN: linezolid; RCLI: Inducible Clindamycin resistance; CF: Fungal Co-infection; CPA: *P. aeruginosa* co-infection; GEN: pts genotype; Esp: sputum; AT: hypopharyngeal suction; MRSA: Methicillin Resistant *S. aureus*; MSSA: Methicillin Sensitive *S. aureus*; SCV: Small colony variant; R: Resistant;S: Susceptible; I: Intermediate; N: Negative; Nt: nontested; Af: *Aspergillus fumigatus*; Ca: *Candida albicans*; Cp: *Candida parapsilosis*; Sp: Scedosporium prolificans; Cd: *Candida dubliniensis*; Pb: *Pseudoallescheria boydii*; Sa: *Scedosporium apiospermum*. X: denotes positive for this feature; -: denotes typical phenotype. See **Table 26** showing the correlation between letter code, CFTR gene mutation of the patient, and bacterial strain isolated from the same patient.

Table 23. The 20 *Pseudomonas aeruginosa* clinical isolates and their characterization by several properties.

GEN	E	E	B	B	B	G	G	U	U	B
CSA							X			
CF				Ca/Cl						
L	X	X		X			X		X	
E			X							
1St					X	X		X		X
COL	S	S	S	S	S	S	S	S	S	S
CEF	R	R	R	R	S	S	S	S	S	S
MB	R	S	S	S	I	I	S	S	S	I
QUIN	R	R	R	R	S	S	S	S	S	S
AM	R	R	R	R	I	S	S	S	S	S
PTC	R	S	S	S	S	S	S	S	S	S
CAR	R	R	R	R	S	S	S	S	S	S
Ph	s	w	sm	s	sm	i	sm	sm	m	i
Str	PA MDR	PA MDR	PA MDR	PA MDR	PA	PA	PA	PA	PA	PA
date	8/8/2006	1/11/2017	6/24/2005	3/27/2017	9/3/2010	8/27/2008	1/31/2017	5/24/2012	9/13/2017	9/6/2010
SAM	ESP	ESP	ESP	ESP	AT	TF	AT	ESP	AT	ESP
ID	21P	22P	23P	24P	25P	26P	27P	28P	29P	30P
ID pt	21	21	22	22	23	24	24	25	25	9

Table 23. The 20 *Pseudomonas aeruginosa* clinical isolates and their characterization by several properties.

GEN	B	F	F	D	D	A	A	B	B	C
CSA	X		X							X
CF			Ca	Ca						Ca
L	X		X		X		X	X	X	
E						X				
1St		X		X						X
COL	S	S	S	S	S	S	S	S	S	S
CEF	S	S	S	S	R	S	R	R	R	S
MB	S	I	I	I	I	I	S	R	I	S
QUIN	R	S	S	CI S/ LE R	R	S	S	R	R	S
AM	S	R	S	S	R	S	R	R	R	S
PTC	S	S	S	S	R	S	S	R	R	S
CAR	S	S	S	MP I/ IP R	R	S	MP S/ IP R	R	R	S
Ph	m	sm	m	i	sm	sm	m	s	m	i
Str	PA	PA	PA	PA	PA MDR	PA	PA	PA MDR	PA MDR	PA
date	1/11/2017	12/5/2006	12/28/2016	5/11/2005	3/29/2017	2/11/2008	2/22/2017	3/7/2006	1/25/2017	7/1/2013
SAM	ESP	AT	AT	ESP	ESP	TF	AT	ESP	ESP	AT
ID	31P	32P	33P	34P	35P	36P	37P	38P	39P	40P
ID pt	9	26	26	27	27	28	28	29	29	30

Table 23. The 20 *Pseudomonas aeruginosa* clinical isolates and their characterization by several properties.

ID pt	ID	SAM	date	Str	Ph	CAR	PTC	AM	QUIN	MB	CEF	COL	1St	E	L	CF	CSA	GEN

ID pt: patient identification;ID: strain code;SAM: Sample; Date: Date of collection; Str: Strain; Ph: Phenotype; CAR: Carbapenems; MP: Meropenem; IP: Imipenem; PTC: Piperacillin/tazobactam; AM: Aminoglycosides; QUIN: Quinolones; CI: Ciprofloxacin; LE: Levofloxacin; MB: Monobactam; CEF: Cephalosporins; COL: Colistin; 1 St: P. aeruginosa first isolate; E: P. aeruginosa early isolate; L: P. aeruginosa late isolate; CF: Fungal co-infection; CSA: S. aureus co-infection; Gen: pts genotype; BP: Biofilm Producer; Esp:sputum; AT: hypopharyngeal suction; TF: throat swabs; PA: P. aeruginosa; PA MDR: P. aeruginosa multi-drug resistant; PA MBL+: P. aeruginosa Metallo-Beta-Lactamases producing; s: small colony phenotype; w- wrinkled colony surface; m: mucoid colony; i: irregular colony edges; sm: smooth phenotype; R: Resistant; S: Susceptible; I: Intermediate; CA: Candida albicans; CL: Candida lusitaniae; X: denotes positive for the feature. See See **Table 26** showing the correlation between letter code, CFTR gene mutation of the patient and bacterial strain isolated from the same patient.

Patients were treated according to current standards of care[158] with at least four microbiological controls per year. Informed consent was obtained from all subjects aged 18 years and older and from parents of all subjects under 18 years of age before enrolment.

Microbiological cultures have been performed according to approved Guidelines, using selective media, manual and automatic systems (API20NE, Vitek2, MALDI-TOF mass spectrometry) for isolates identification and 16S rRNA sequencing to assess ambiguous identifications. The strains were selected from a local bacteria collection, including about 10.000 CF bacterial isolates. The species *S. aureus* and *P. aeruginosa* have been chosen for their clinical relevance in CF disease's natural history since they are related to a worst prognostic impact compared to other pathogens whose role is still under discussion.

In order to represent the complexity of the CF lung microbiota population attending OPBG Center, a selection of specific strains with different phenotypic and biochemical features has been performed. The strains' characteristics are described in (**Table 22** and **Table 23**).

5.2.3 Qualitative description of the clinical isolates

Twenty *S. aureus* strains with a different susceptibility profile, belonging to 20 CF patients, were selected: 10 Methicillin-Sensitive (MSSA) and 10 Methicillin-Resistant (MRSA). Among the MRSA strains, three *S. aureus* with phenotypic "small colony variants" (SCVs) have been chosen, characterized by the slow growth of small, unpigmented, non-hemolytic colonies.

Antimicrobial susceptibility profiles of MSSA and MRSA isolates were defined by automatic system Vitek2 (Biomerieux, France) or manual system E-test (Liofilchem, Italy). In particular, susceptibility to quinolones (ciprofloxacin, levofloxacin), trimethoprim/sulfamethoxazole, erythromycin, clindamycin, linezolid was assessed, according to EUCAST (www. EUCAST.org) criteria. Moreover, the clindamycin-inducing resistance test (40% positive test) was performed to classify *S. aureus* isolates that could develop acquired resistance to erythromycin or other macrolides during therapy with this antibiotic (**Table 22**)[159].

Twenty *P. aeruginosa* isolates belonging to 11 CF patients were also selected (**Table 23**). The selected strains had been categorized as first, early, and late isolates. In particular, seven strains have been associated with the first acquisition of *P. aeruginosa* (first strains), two strains have been isolated one year after the first acquisition (early strains), and 11 strains have been isolated at least five years after the onset of chronic colonization (late strains).

Moreover, different phenotypes (mucoid, wrinkle surface, irregular edges, or smooth) and strains with different antibiotic susceptibility patterns, e.g., *P. aeruginosa* producing Metallo-Beta-Lactamases (MBL)[160] or *P. aeruginosa* multi-drug resistant (MDR), have been selected.

According to EUCAST criteria, susceptibility testing to carbapenems (imipenem, meropenem), piperacillin/tazobactam, aminoglycosides (tobramycin, amikacin), quinolones (ciprofloxacin, levofloxacin), monobactam (aztreonam), and cephalosporins (ceftazidime, cefepime) was carried out by Minimum Inhibitory Concentration (MIC) determined by E-test on Mueller Hinton (MH) agar plates. The colistin MIC values were evaluated by Broth Microdilution (ComASP Colistin Liofilchem, Italy); 35% of *P. aeruginosa* isolates were MDR (i.e., resistant to three or more classes of antimicrobials)[161] (**Table 24**). **Table 25** reports that the percentage of bacterial strains resulted sensitive or resistant to different classes of antibiotics here tested. Co-infection by bacterial (*P. aeruginosa* and *S. aureus*) and fungal agents (*Aspergillus fumigatus*, *Candida albicans*, *Candida parapsilo*sis, *Candida dubliniensis*, *Candida lusitaniae*, *Scedosporium prolificans*, *Scedosporium apiospermum*, *Pseudoallescheria boydii*) was also evaluated for each patient (**Table 22** and **Table 23**). **Table 26** reports letters' code correspondence for the strains associated genotype reported in **Table 22** and **Table 23**.

Table 24. Antimicrobial activity corresponding to minimal bactericidal concentration of previously selected EOs on all 40 clinical isolates.

Bacterial strains	CEO	BEO	CCPEO	Bacterial strains	CEO	BEO	CCPEO
ATCC6538P	1%	1%	1%	PA O1	1%	1%	1%
ATCC25923	1%	1%	1%	PA 14	1%	1%	1%
SA01	1%	1%	1%	PA21	1%	1%	1%
SA02	1%	1%	1%	PA22	1%	1%	1%
SA03	1%	1%	1%	PA23	1%	1%	1%
SA04	1%	1%	1%	PA24	1%	1%	1%
SA05	1%	1%	0.1%	PA25	1%	1%	1%
SA06	1%	1%	1%	PA26	1%	1%	1%
SA07	1%	1%	1%	PA27	1%	1%	1%
SA08	1%	1%	1%	PA28	1%	1%	1%
SA09	1%	1%	1%	PA29	1%	1%	1%
SA10	1%	1%	1%	PA30	1%	1%	1%
SA11	1%	1%	1%	PA31	1%	1%	1%
SA12	1%	1%	1%	PA32	1%	1%	1%
SA13	1%	1%	1%	PA33	1%	1%	1%
SA14	1%	1%	1%	PA34	1%	1%	1%
SA15	1%	1%	1%	PA35	1%	1%	1%
SA16	1%	1%	1%	PA36	1%	1%	1%
SA17	1%	1%	1%	PA37	1%	1%	1%
SA18	1%	1%	1%	PA38	1%	1%	1%
SA19	1%	1%	1%	PA39	1%	1%	1%
SA20	1%	1%	1%	PA40	1%	1%	1%

Table 25. Percentage of susceptibility of *S. aureus* and *P. aeruginosa*

Classes of antimicrobials	Molecules	Susceptibility according to EUCAST[a]					
		S.aureus			***P.aeruginosa***		
		% S	% I	% R	% S	% I	% R
Quinolones	CI	60	5	35	60	-	40
	LE	60	5	35	55	-	45
Sulfonamides	T/S	100	-	-		Nt	
Macrolides	ER	10	10	80		Nt	
Lincosamides	CLI	25	5	70		Nt	
Oxazolidinones	LIN	100	-	-		Nt	
Carbapenems	IP		Nt		55	-	45
	MP		Nt		60	5	35
Penicillins + β lactamase inhibitors	PTC		Nt		80	-	20
Aminoglycosides	TM		Nt		50	5	45
	AK		Nt		50	5	45
Monobactams	AT		Nt		45	45	10
Cephalosporins	CAZ		Nt		60	-	40
	PM		Nt		60	-	40
Polymyxins	CO		Nt		100	-	-

Table 25. Percentage of susceptibility of *S. aureus* and *P. aeruginosa*

S, susceptible; I, intermediate; R, resistant; CI, ciprofloxacin; LE, levofloxacin; T/S, trimethoprim/sulfamethoxazole; ER, erythromycin; CLI, clindamycin; LIN, linezolid; IP, imipenem; MP, meropenem; PTC, piperacillin-tazobactam; TM, tobramycin; AK, amikacin; AT, aztreonam; CAZ, ceftazidime; PM, cefepime, CO, colistin. Nt, non-tested.

[a] EUCAST breakpoints (in milligrams per liter) for *S.aureus*: CI-5µg , S ≤ 1 - R > 1; LE-5µg , S ≤ 1 - R > 1;T/S - 1.25/23.75 µg , S ≤ 4 - R > 4; ER-15 µg , S ≤ 1 - R > 2; CLI-2 µg , S ≤ 0.25 - R > 0.5; LIN-10 µg , S ≤ 4 - R > 4;

[a] EUCAST breakpoints (in milligrams per liter) for *P.aeruginosa*: CI-5µg , S ≤ 0.5 - R > 0.5; LE-5µg , S ≤ 1 - R > 1; IP-10 µg , S ≤ 4 - R > 8; MP-10 µg , S ≤ 2 - R > 8; PTC –100/10 µg, S ≤ 16 - R > 16; TM –10 µg, S ≤ 4 - R > 4; AK–30 µg, S ≤ 8 - R > 16; AT–30 µg, S ≤ 1 - R > 16; CAZ–30 µg, S ≤ 8 - R > 8; PM–30 µg, S ≤ 8 - R > 8; CO–10 µg, S ≤ 2 - R > 2.

Table 26. Table shows the correlation between letter code, CFTR gene mutation of the patient and bacterial strain isolated from the same patient.

Code	Genotype	ID strain
A	621+1G > T/R553X	4S
B	F508del/1717-1G- > A	13S
C	F508del/F508del	5S
C	F508del/F508del	9S
C	F508del/F508del	11S
C	F508del/F508del	14S
C	F508del/F508del	15S
C	F508del/F508del	18S
D	F508del/G1244E	7S
E	F508del/G542X	3S
F	F508del/L1077P	20S
G	F508del/R1162X	10S
H	F508del/R117L + L997F	16S
I	F508del/R585X	8S
J	F508del/W1282X	1S
K	G542X/3271 + 42A/T	6S
L	L636P/P499A	19S
M	N1303K/2184insA	17S
N	Q220X/A1006E	2S
U	None	12S
A	F508del/E193K	36P
A	F508del/E193K	37P
B	F508del/F508del	23P
B	F508del/F508del	24P
B	F508del/F508del	25P
B	F508del/F508del	30P
B	F508del/F508del	31P
B	F508del/F508del	38P
B	F508del/F508del	39P
C	F508del/G542X	40P
D	F508del/I1234V	34P
D	F508del/I1234V	35P
E	N1303K/3849 + 10kbC > T	21P
E	N1303K/3849 + 10kbC > T	22P
F	R347P/L571S	32P
F	R347P/L571S	33P
G	W1282X/2789 + 5G- > A	26P
G	W1282X/2789 + 5G- > A	27P
U	None	28P
U	None	29P

5.2.4 Biofilm production assay

The quantification of biofilm production was based on microtiter plate biofilm assay (MTP), as reported in literature[93]. Briefly, the wells of a sterile 96-well flat-bottomed polystyrene plate were filled with 100 µL of the appropriate medium. 1/100 dilution of overnight bacterial cultures were added into each well (about 0.5 OD 600 nm). The plates were incubated aerobically for 18 hours at 37 °C. Biofilm formation was measured using crystal violet staining. After incubation, planktonic cells were gently removed; each well was washed three times with double-distilled water and patted dry with a piece of paper towel in an inverted position. Each well was stained with 0.1% crystal violet and incubated for 15 minutes at room temperature, rinsed twice with double-distilled water, and thoroughly dried to quantify biofilm formation. The dye bound to adherent cells was solubilized with 20% (v/v) glacial acetic acid and 80% (v/v) ethanol. After 30 min of incubation at room temperature, OD590 was measured to quantify the biofilm's total biomass formed in each well. Each data point is composed of 4 independent experiments, each performed at least in 6-replicates.

5.2.5 Statistical analysis of biological evaluation

Data reported were statistically validated using Student's *t*-test comparing mean absorbance of treated and untreated samples. The significance of differences between mean absorbance values was calculated using a two-tailed Student's *t*-test. A p-value of <0.05 was considered significant.

5.2.6 Chemical composition analysis of active selected essential oils

EOs were purchased from Farmalabor srl (Assago, Italy) and analyzed to characterize their composition. Chemical analyses of EOs were performed by a Turbomass Clarus 500 GC-MS/GC-FID from Perkin Elmer instruments (Waltham, MA, USA) equipped with a Stabilwax fused-silica capillary column (Restek, Bellefonte, PA, USA) (60 m × 0.25 mm, 0.25 mm film thickness). The operating conditions used were as follows: GC oven temperature was kept at 40 °C for 5 min and programmed to 220 °C at a rate of 6 °C/minute and kept constant at 220 °C for 20 minutes. Helium was used as carrier gas (1.0 mL/min). Solvent delay 0–2 min and scan time 0.2 seconds. The mass range was from 30 to 350 m/z using electron-impact at 70 eV mode. 1 µL of each essential oil was diluted in 1 mL of methanol and 1 µL of the solution was injected into the GC injector at the temperature of 280 °C. Relative percentages for quantification of the components were calculated by the electronic integration of the GC-FID peak areas. The constituents' identification was achieved by comparing the obtained mass spectra for each component with those reported in mass spectra Nist and Wiley libraries. Linear retention indices (LRI) of each compound were calculated using a mixture of aliphatic hydrocarbons (C8-C30, Ultrasci) injected directly into the GC injector at the same temperature program reported above.

5.2.7 Determination of EOs minimal inhibitory concentration (MIC)

The MIC was determined as the lowest concentration at which the observable bacterial growth was inhibited. MICs were determined according to the guidelines of Clinical Laboratory Standards Institute[127] (CLSI). Each EO was solubilized by adding DMSO to generate a 1 g/mL mother stock solution. Appropriate dilution (10^6 cfu/mL) of bacterial culture in the exponential phase was used. The antimicrobial activity of each EO was evaluated at a concentration of 1 mg/mL range. Experiments were performed in quadruplicate.

5.2.8 Unsupervised machine learning clusterization of clinical isolates

The cluster analysis was implemented in the Python (version 3.6) programming language. The *S. aureus* and *P. aeruginosa* datasets were imported in a Jupyter-notebook[129] (version 5.7), and the categorical variables loaded into a Pandas35 data frame were transformed into dummy indicator variables for the subsequent Principal Component Analysis (PCA) using the utilities available in the Pandas[70,71] (version 0.23) library. The PCA analysis was performed using the scikit-learn library[72] (version 0.20) to extract the first 20 principal components (**Figure 32**). The scores and loadings were graphically inspected on plots generated using the matplotlib library[69](version 3.0) (**Figure 33**). The PCs were used as features for the k-means clusterization. Silhouette analysis[162] was performed to evaluate the separation distance between the resulting clusters and choose an optimal value for the cluster number. The optimal number of clusters was identified by the maximum silhouette scores, as graphically reported in **Figure 34**. Through k-means, the centroid of each cluster was calculated, and the closest data point directly indicated the RS (**Figure 35**).

5.3 Results

5.3.1 Characterization of biofilm formation of clinical bacterial strains

Clinical bacterial strains were investigated for their ability to produce biofilm (**Figure 31**). Biofilm formation was evaluated at 37 °C in BHI for 18 hours, as described in the Materials and Methods section. Biofilm formation was also evaluated for four reference strains included in the experimental plan. **Figure 31** (top panel) reports biofilm formation of bacterial strains belonging to *S. aureus* species. Clinical strains, named from 1S to 20S, were classified as "weak" or "moderate" biofilm producers, according to Cafiso and coworker, 2007[145]. Both reference strains for *S. aureus* species are strong biofilm producers. **Figure 31** (bottom panel) reports biofilm formation of bacterial strains belonging to *P. aeruginosa* species. Clinical strains, named from 21P to 40P, were classified as: "non producer", "weak", "moderate" and "strong" biofilm producers, according to Perez and Barth, 2011[156] (**Table 21**).

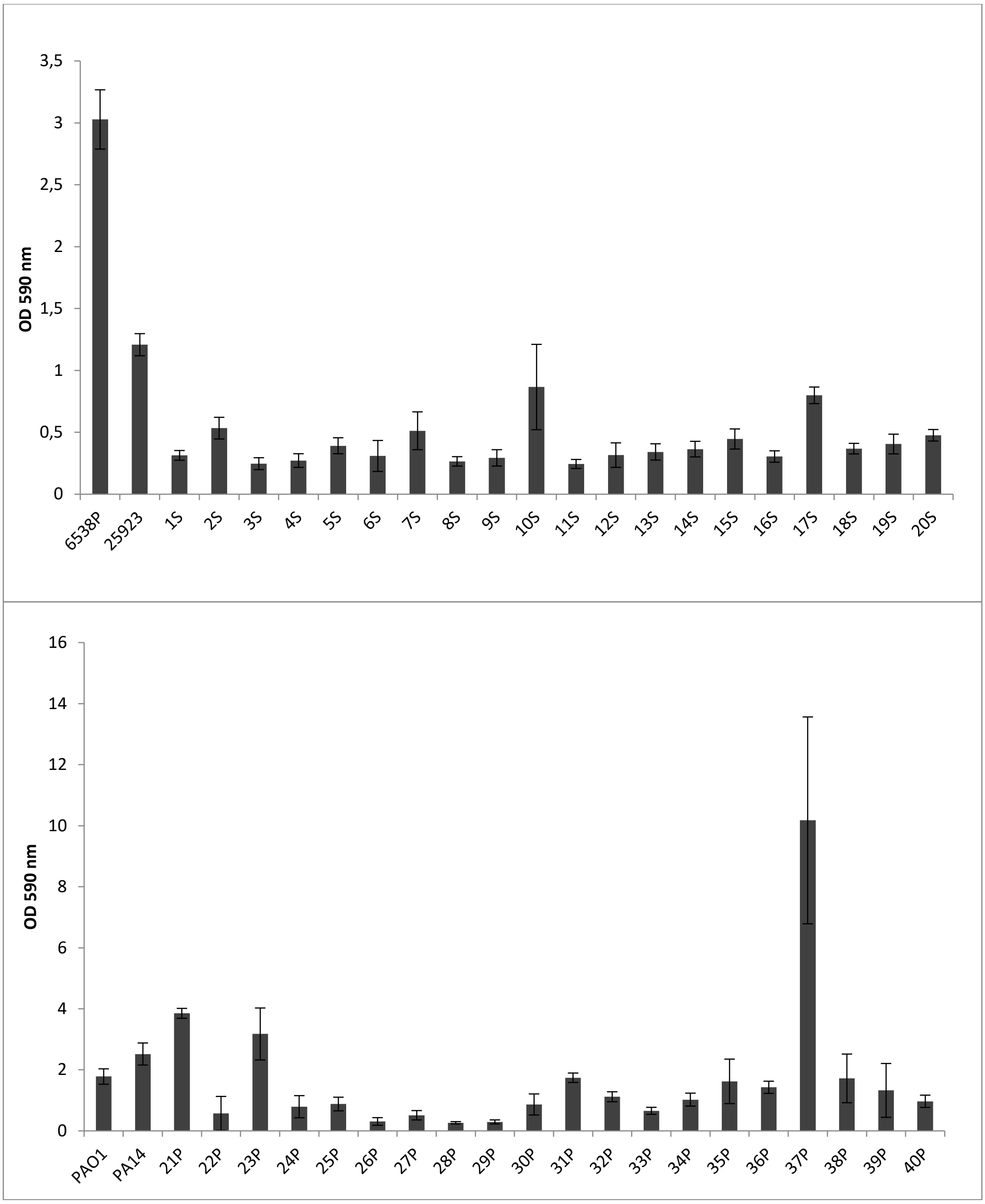


Figure 31. Biofilm formation of *S. aureus* clinical and reference strains (top panel) and *P. aeruginosa* clinical and reference strains (bottom panel). The biofilm formation was evaluated after 18 hours incubation in polystyrene plates at 37 °C. The data are reported as OD 590 nm after crystal violet staining. Each data point represents the mean ± SD of four independent samples.

5.3.2 Selection of representative microorganisms by machine learning

The 40 selected strains were divided according to the main strains families into S. aureus and P. aeruginosa datasets and imported into a python pandas data frame. The principal components analysis (PCA) indicated that 90% of the variance is explained by the first 10th principal components (PCs) (**Figure 32**). Nevertheless, visual inspection of the first principal component (PC1) versus the second principal component (PC2) scores and loadings plots indicated the PCs as potential new variables to cluster the datasets (**Figure 33**). Application of the Silhouette Analysis[162] coupled with the k-means clustering[24] to the first two PCs indicated the optimal number of clusters to be 6 and 3 for the *P. aeruginosa* and *S. aureus* strains, respectively (**Figure 34** and **Figure 35**).

The nearest data point to cluster centroid was selected for each cluster, yielding to a selection of representative strains to be screened with the commercial EOs. Analysis of data revealed the six samples, precisely 22P, 25P, 26P, 27P, 37P, and 39P as the representatives for *P. aeruginosa*, whereas samples 4S, 5S, and 19S were selected for *S. aureus*.

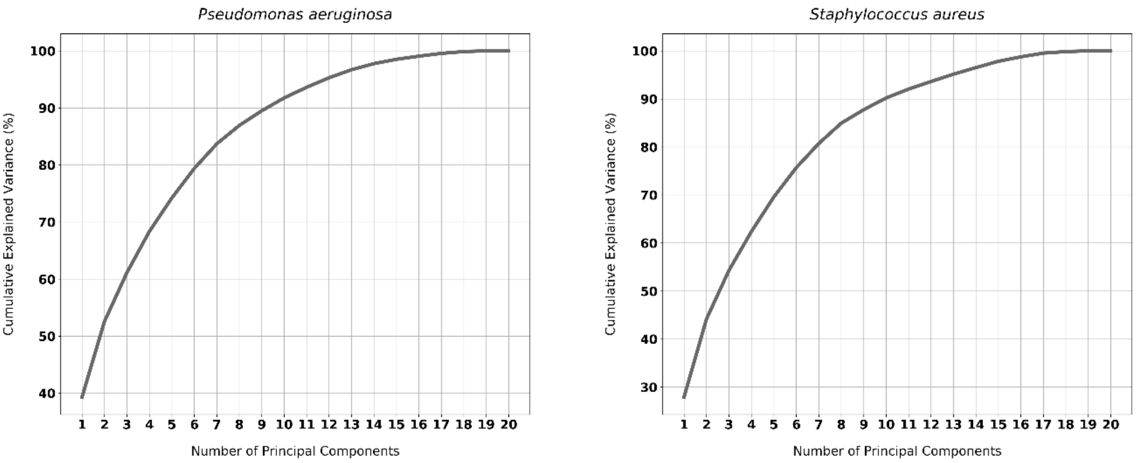


Figure 32. Explained cumulative variance versus the number of extracted PCs for the PA (left panel) and SA (right panel) datasets

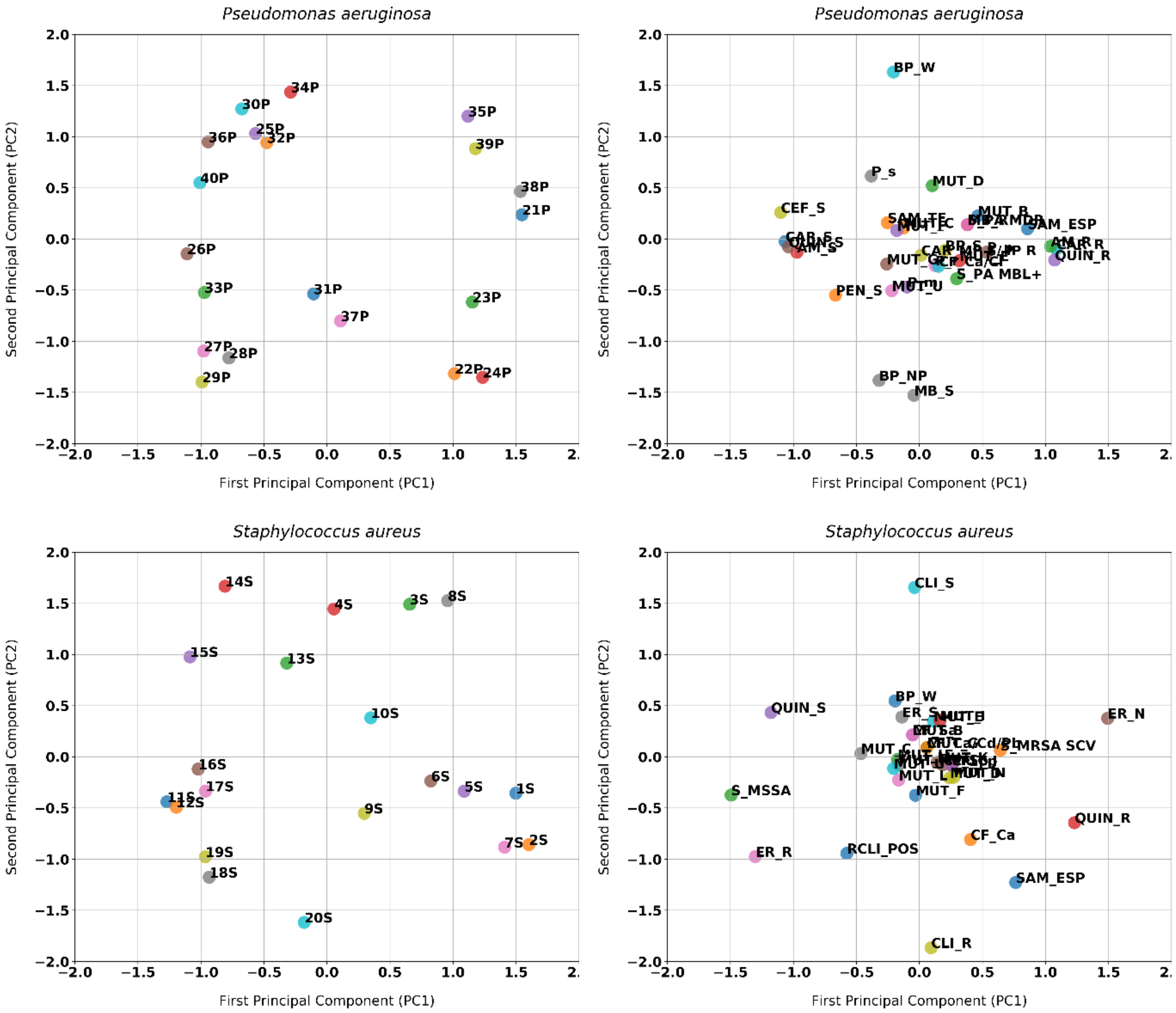


Figure 33. Scores (left panels) and loadings (right panels) plots for the *Pseudomonas aeruginosa* (top panels) and *the Staphylococcus aureus* (bottom panels).

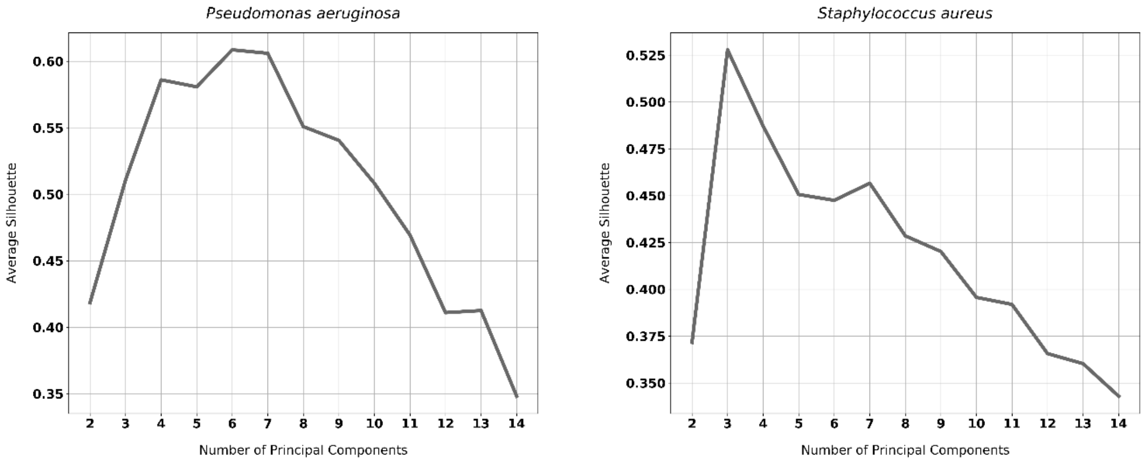


Figure 34. Average silhouette scores versus the number of clusters. The maximum indicates the optimal number of clusters.

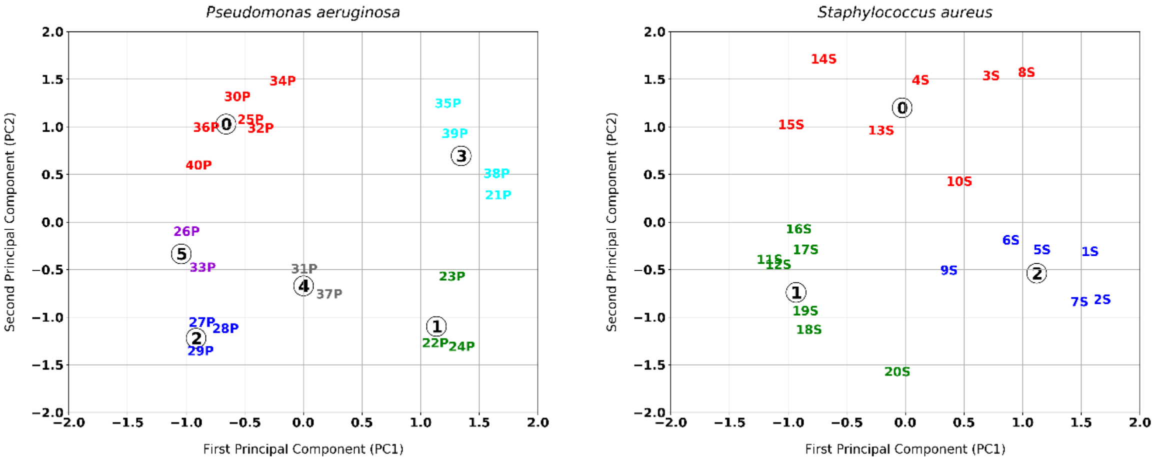


Figure 35. Score plots indicating the clustered *Pseudomonas aeruginosa* (left panel) and *Staphylococcus aureus* (right panel) datasets. The kmeans centroids of each cluster are also reported.

5.3.3 Antimicrobial activity of EOs on *P. aeruginosa* and *S. aureus* clinical strains from cystic fibrosis patients

Essential oils were tested for their ability to inhibit bacterial growth of *P. aeruginosa* and *S. aureus* clinical and reference strains. Analysis was performed on three representative *S. aureus* strains and six representative *P. aeruginosa* strains, previously selected by machine learning analysis. EOs were tested at a concentration of 1% v/v (). Several EOs have shown antimicrobial activity on many bacterial strains. It is worthy to note that the *P. aeruginosa* reference strain PAO1 is the most resistant to the action of EOs, since only four EOs inhibited it. This analysis allowed to identify three EOs active against all the representative strains used, namely cade essential oil (22 in **Table 28**, CEO), birch essential oil (32 in **Table 28**, BEO) and Ceylon cinnamon peel essential oil (39 in **Table 28**, CCPEO). Thus, these 22, 32 and 39 EOs were tested on all clinical bacterial strains. Results summarized in **Table 24** confirmed that BEO, CEO and CCPEO exerted a strong and effective bactericidal potency on all tested clinical strains.

Table 27. Antimicrobial activity of EOs listed in **Table 28** on representative clinical strains and reference strains of *S. aureus* and *P. aeruginosa*.

Eos ID	6538P	25923	4S	5S	19S	PaO1	PA14	22P	25P	26P	27P	37P	39P
1	1%	1%	1%										
2													
3								1%				1%	
4	1%	1%	1%	1%	1%			1%				1%	
5													
6		1%	1%	1%	1%								
7													
8													
9	1%	1%	1%	1%	1%		1%	1%	1%	1%		1%	1%
10													
11													
12													
13													
14	1%	1%	1%	1%	1%	1%	1%		1%		1%	1%	1%
15													
16													
17													
18													
19													
20													

Table 27. Antimicrobial activity of EOs listed in **Table 28** on representative clinical strains and reference strains of *S. aureus* and *P. aeruginosa*.

Eos ID	6538P	25923	4S	5S	19S	PaO1	PA14	22P	25P	26P	27P	37P	39P
21													
22	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
23													
24													
25													
26													
27													
28													
29													
30													
31													
32	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
33													
34													
35													
36													
37	1%	1%	1%	1%	1%			1%				1%	1%
38												1%	
39	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
40													
41													
42													
43													
44													
45													
46	1%	1%	1%	1%	1%							1%	1%
47													
48	1%	1%	1%	1%	1%		1%						1%
49													

Table 27. Antimicrobial activity of EOs listed in **Table 28** on representative clinical strains and reference strains of *S. aureus* and *P. aeruginosa*.

Eos ID	6538P	25923	4S	5S	19S	PaO1	PA14	22P	25P	26P	27P	37P	39P
50													
51													
52				1%	1%								
53													
54	1%				1%								
55													
56													
57													
58													
59	1%	1%	1%	1%	1%							1%	1%
60				1%	1%							1%	1%
61	1%			1%									1%

Table 28. Essential oils names.

EO ID	EO Name	EO ID	EO Name
1	Chamomile Morocco Essential Oil	32	Birch Essential Oil
2	Sage Sclarea Essential Oil	33	Fennel Essential Oil
3	Salvia Officinalis Essential Oil	34	Cedar Fruit Essential Oil
4	Red Thyme Essential Oil	35	Lemon Essential Oil
5	Tea Tree Oil	36	Roman Chamomile Essential Oil
6	Melissa Oiio Essential	37	Savory Essential Oil
7	Pinus Mugo Essential Oil	38	Rosemary Essential Oil
8	Geranium Bourbon Essential Oil	39	Ceylon Cinnamon Peel Essential Oil
9	Oregano Essential Oil	40	Eucaliptus Globulus Essential Oil
10	Ylang Ylang Essential Oil	41	Sweet Orange Essential Oil
11	Coriander Essential Oil	42	Niaouly Essential Oil
12	Lavandula Angustifoglia Essential Oil	43	Artemisia Essential Oil
13	Myrtle Essential Oil	44	Cajeput Essential Oil
14	Garlic Essential Oil	45	Black Pepper Essential Oil
15	Cardamom Essential Oil	46	White Thyme Essential Oil
16	Mandarin Essential Oil	47	Marjoram Essential Oil
17	Hyssop Essential Oil	48	Cloves Essential Oil
18	Grapefruit Essential Oil	49	Cypress Essential Oil
19	Cymbopogon Essential Oil	50	Nutmeg Natural Essential Oil
20	Pinus Sibirica Essential Oil	51	Peppermint Essential Oil
21	Camphor Essential Oil	52	Verbena officinalis Essential Oil
22	Cadè Essential Oil	53	Basil Essential Oil
23	Cedar Leaves Essential Oil	54	Cymbopogon martinii Essential Oil
24	Ginger Essential Oil	55	Laurel Essential Oil

25	Cumin Essential Oil	56	Anise Essential Oil
26	Patchouli Essential Oil	57	Incense Essential Oil
27	Bitter Orange Essential Oil	58	Mentha Suaveolens (Sicily) Essential Oil
28	Eucalyptus Essential Oil	59	Coridotthymus Capitatus (Sicily) Essential Oil
29	Pinus Silvester Essential Oil	60	Thymus Vulgaris (Sicily) Essential Oil
30	Bergamot Essential Oil	61	Origanum Hirtum (Sicily) Essential Oil
31	Juniper Essential Oil		

5.3.4 Chemical composition analysis of active selected essential oils

The results of GC and GC-MS analyses of the essential oils are reported in **Table 29**, **Table 30** and **Table 31**. In the BEO, 21 components were identified and the major constituents were δ-cadinene, calamenene and creosol (22.2%, 15.2% and 12.8% respectively) (**Table 29**). The chemical composition of CCPEO was characterized by the presence of 19 compounds and by a high amount of cinnamaldehyde (49.4%) followed by eugenol (21.2%) (**Table 30**). The chemical composition of the CEO indicated 21 components and the most abundant were delta-cadinene (27.7%), calamenene (14.8%) and creosol (12.6%) (**Table 31**). At first glance, the CEO's chemical composition seems very similar to that of BEO as the main compounds showed comparable percentages. Among the minor components of CEO, α-selinene (2.2%), aromadendrene (1.1%), and gleenol (1.1%) were found, whereas isoledene (5.7%) was found in BEO. At a deeper analysis, the qualitative chemical profiles were compared, and a 0.62 Tanimoto index was calculated, thus indicating that although displaying a similar chromatogram, the two EOs are indeed different. EOs producer was also inquired, and their technical staff confirmed the two oils were sharing high similarity quantitative profile in the main constituents.

Table 29. Chemical composition (%) of Birch EO.

#[1]	Name	RI[2]	RI^{lit3}	Area %
1	2-cyclopenten-1-one, 3-methyl	1510	1513	0.5
2	2-cyclopenten-1-one, 2,3-dimethyl	1528	1535	0.7
3	α-cedrene	1590	1599	9.9
4	dihydrocurcumene	1610	*	3.5
5	isoledene	1655	*	5.7
6	α-muurolene	1685	1690	4.4
7	δ-cadinene	1758	1758	22.2
8	calamenene	1802	1804	15.2
9	α-methylnaphtalene	1889	1891	0.9
10	guaiacol	1895	1897	3.5
11	isolongifolene, 4,5,9,10-dehydro-	1920	*	1.8
12	creosol	1960	1956	12.8
13	o-creosol	2009	2011	0.9
14	phenol	2011	2012	1.1
15	p-ethylguaiacol	2027	2032	5.0
16	m-cresol	2075	2081	1.4
17	phenol, 2,5-dimethyl	2080	2085	1.8
18	p-propylguaiacol	2102	2103	2.1
19	eugenol	2155	2166	0.9
20	cadalene	2199	2200	4.6
21	isoeugenol	2270	2268	0.9
	Unidentified compounds			0.2

[1] Compound identification number.

[2] Retention indices measured on polar column.

[3] Retention indices from literature;

*RI^{lit} not available for polar column.

$^{+}$ Normal alkane RI.

Table 30. Chemical composition (%) of Ceylon Cinnamon EO.

#[1]	Name	RI[2]	RI^{lit3}	Area %
1	α-pinene	1021	1021	0.9
2	camphene	1060	1065	0.3
3	β-pinene	1100	1105	0.3
4	α-phellandrene	1161	1160	0.9
5	limonene	1200	1198	0.5
6	β-phellandrene	1203	1204	1.4
7	m-cymene	1260	1258	2.1
8	α-copaene	1490	1487	0.4
9	β-linalool	1535	1537	5.2
10	terpinen-4-ol	1600	1603	0.3
11	β-caryophyllene	1617	1619	4.7
12	cinnamaldheyde, o-methoxy-	1650	*	1.0
13	humulene	1665	1668	0.8
14	α-terpineol	1677	1675	0.4
15	safrole	1870	1874	0.3
16	cinnamaldheyde	2037	2049	49.4
17	eugenol	2170	2175	21.2
18	eugenol acetate	2270	2277^{+}	0.9
19	benzyl benzoate	2648	2652	8.6
	Unidentified compounds			0.4

[1] Compound identification number.

[2] Retention indices measured on polar column.

[3] Retention indices from literature;

*RI^{lit} not available for polar column.

$^{+}$ Normal alkane RI.

Table 31. Chemical composition (%) of Cadè EO.

#[1]	Name	RI[2]	RI^lit3	Area %
1	2-cyclopenten-1-one, 3-methyl	1510	1513	0.7
2	α-cedrene	1590	1599	7.9
3	aromadendrene	1609	1610	1.1
4	dihydrocurcumene	1690	1696	3.5
5	α-selinene	1751	1750	2.2
6	α-muurolene	1755	*	4.0
7	δ-cadinene	1758	1758	27.7
8	calamenene	1802	1832	14.8
9	phenol, 2-methoxy-	1838	1846	4.7
10	isolongifolene, 4,5,9,10-dehydro-	1920	*	1.7
11	creosol	1960	1956	12.6
12	o-creosol	2009	2011	1.0
13	phenol	2011	2012	1.0
14	p-ethylguaiacol	2027	2032	5.4
15	m-cresol	2075	2081	1.5
16	phenol, 3-methyl-	2095	2099	1.1
17	p-propylguaiacol	2102	2103	1.9
18	eugenol	2155	2175	0.4
19	gleenol	2175	*	1.1
20	cadalene	2199	2200	4.3
21	isoeugenol	2270	2268	1.2
	Unidentified compounds			0.2

[1] Compound identification number.

[2] Retention indices measured on polar column.

[3] Retention indices from literature;

*RI^{lit} not available for polar column.

$^{+}$ Normal alkane RI.

5.4 Discussion and Conclusions

Long-term antibiotic administration to prevent and treat airway infections in CF patients has been shown to be associated with the emergence of multi-drug (MDR) antimicrobial-resistant microorganisms[163]. In particular, *mecA/mecC* genes acquisition in *S. aureus* and accumulation of resistance mechanisms after antibiotic exposure in *P. aeruginosa*, both key pathogens in CF lung, are a concern in this context[164,165]. Multidrug resistance significantly limits effective therapeutic options, affecting the clinical outcome and prognosis of patients. For this reason, the identification and development of new antibacterial agents are fundamental to improve the survival and quality of life of individuals with CF. Therefore the development of antimicrobial agents provided with novel molecular mechanisms that may control bacterial infectious diseases without diffusing antibacterial resistance is desirable[166]. Unsupervised Machine Learning algorithms applied to a panel of 40 strains of *S. aureus* and *P. aeruginosa* isolated from CF patients led to select fewer representative strains using phenotypical and genotypical characteristics as categorical descriptors. Therefore, the antibacterial activity of all tested EOs was initially assessed on nine selected bacterial strains: six representative strains for *P. aeruginosa* and three representative strains for *S. aureus*. The activity of all 61 EOs was also assessed on reference strains. Antimicrobial assays led to identifying 3 EOs (CEO, BEO, and CCPEO) out of the tested 61, which exhibited the highest antibacterial activity on the previously selected bacterial strains and reference ones. The antibacterial activity of the three selected EOs was then extended to all strains of both species.

Interestingly all three EOs showed the utmost antimicrobial potency on all studied strains. Nothing can be yet ruled out on the chemical compounds' role. Future studies involving machine learning applications will be dedicated to investigating the importance of chemical constituent either on biofilm modulation or in antibacterial potencies—several papers aimed at elucidating the antimicrobial mechanism of action of EOs. For example, cinnamaldehyde, the primary component of cinnamon, can disrupt the transmembrane potential of *P. aeruginosa*[167].

Furthermore, EOs of different origins (lavender, lemongrass, marjoram, peppermint, tea tree and rosewood) show antimicrobial activity against Burkholderia cepacia complex inducing changes in membrane fatty acid composition, followed by membrane disruption[168]. Also, EO from *Alluaudia procera* was active against *S. aureus* ATCC25923, a multi-resistant strain[169].

Reported data confirmed the possibility of using EOs as therapeutic strategies in multi-resistant strains, probably due to the heterogeneous composition of the oils themselves. Notably, in this work, we found EOs antibacterial activity unrelated to each strain's antibiotic resistance profile. This observation is relevant as it suggests the EOs potential uses by topical administration without considering the complexity of the microbiota's drug resistance profile in every patient.

In conclusion, the approach herein applied allowed to minimize the experimental steps, and it was possible to identify the most promising EOs based on probabilistic evaluations that confirmed their broad spectra of antibacterial potency with a reduced set of experiments.

From a literature survey (www.scopus.com, accessed 2019 December 13, keywords: essential oil, antibacterial activity, and resistance), no evidence of resistance to EOs antibacterial activity has yet been reported. This is a characteristic particularly relevant for antibacterial candidates to be administered for a chronic disease such as CF. Indeed some papers report an increase of susceptibility to antibiotics after treatment with essential oils[170,171]. Although a plethora of publications did not show the development of resistance to EOs, a very recent publication suggested the induction of efflux pumps and multidrug resistance in *P. aeruginosa* by Cinnamaldehyde, the main component of cinnamon[172]. Therefore, considering the recent reports, much still needs to be clarified on the essential oils' effect on bacterial multi-drug resistance.

6 References

1. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **3**, 210–229 (1959).

2. Mitchell, T. M. *Machine Learning, McGraw-Hill Higher Education*. *New York* (1997).

3. Alzubi, J., Nayyar, A. & Kumar, A. Machine Learning from Theory to Algorithms: An Overview. *J. Phys. Conf. Ser.* **1142**, 12012 (2018).

4. Russell, S. & Norvig, P. *Artificial Intelligence A Modern Approach Third Edition*. *Pearson* (2010). doi:10.1017/S0269888900007724.

5. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **9**, 1–10 (2019).

6. Mucherino, A., Papajorgji, P. J. & Pardalos, P. M. k-Nearest Neighbor Classification. in 83–106 (Springer, New York, NY, 2009). doi:10.1007/978-0-387-88615-2_4.

7. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* (1995) doi:10.1023/A:1022627411411.

8. Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* (1998) doi:10.1109/34.709601.

9. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* (2001) doi:10.2307/2699986.

10. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).

11. Tropsha, A., Gramatica, P. & Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. in *QSAR and Combinatorial Science* vol. 22 69–77 (Wiley-VCH Verlag, 2003).

12. Rücker, C., Rücker, G. & Meringer, M. Y-randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **47**, 2345–2357 (2007).

13. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. *Algorithms for Hyper-Parameter Optimization*.

14. Bergstra, J., Ca, J. B. & Ca, Y. B. *Random Search for Hyper-Parameter Optimization*

Yoshua Bengio. *Journal of Machine Learning Research* vol. 13 http://scikit-learn.sourceforge.net. (2012).

15. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. in *Advances in Neural Information Processing Systems* (2012).

16. Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.* (1978) doi:10.1016/S0001-2998(78)80014-2.

17. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct.* (1975) doi:10.1016/0005-2795(75)90109-9.

18. Box, P. O., Van Der Maaten, L., Postma, E. & Van Den Herik, J. *Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review*. http://www.uvt.nl/ticc (2009).

19. Pudil, P. & Novovičová, J. Novel Methods for Feature Subset Selection with Respect to Problem Knowledge. in *Feature Extraction, Construction and Selection* 101–116 (Springer US, 1998). doi:10.1007/978-1-4615-5725-8_7.

20. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (1999) doi:10.1111/1467-9868.00196.

21. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

22. Szekely, G. J. & Rizzo, M. L. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J. Classif.* **22**, 151–183 (2005).

23. Andreopoulos, B., An, A., Wang, X. & Schroeder, M. A roadmap of clustering algorithms: Finding a match for a biomedical application. *Briefings in Bioinformatics* vol. 10 297–314 (2009).

24. Celebi, M. E., Kingravi, H. A. & Vela, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst. Appl.* (2013) doi:10.1016/j.eswa.2012.07.021.

25. Sugiyama, M. Maximum Likelihood Estimation for Gaussian Mixture Model. in *Introduction to Statistical Machine Learning* 157–168 (Elsevier, 2016). doi:10.1016/B978-0-12-802121-7.00026-1.

26. Ester, M., Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. 226--231 (1996).

27. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 1–21 (2017).

28. Muratov, E. N. *et al.* QSAR without borders. *Chem. Soc. Rev.* **49**, 3525–3564 (2020).

29. Cherkasov, A. *et al.* QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry* vol. 57 4977–5010 (2014).

30. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 45 (2017).

31. Sheridan, R. *et al.* Toward structure-based predictive tools for the selection of chiral stationary phases for the chromatographic separation of enantiomers. *J. Chromatogr. A* **1467**, 206–213 (2016).

32. Simón-Vidal, L. *et al.* Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *J. Chem. Inf. Model.* **58**, 1384–1396 (2018).

33. Ban, F. *et al.* Best Practices of Computer-Aided Drug Discovery: Lessons Learned from the Development of a Preclinical Candidate for Prostate Cancer with a New Mechanism of Action. *Journal of Chemical Information and Modeling* vol. 57 1018–1028 (2017).

34. Patsilinakos, A., Ragno, R., Carradori, S., Petralito, S. & Cesa, S. Carotenoid content of Goji berries: CIELAB, HPLC-DAD analyses and quantitative correlation. *Food Chem.* **268**, (2018).

35. Božovic, M. *et al.* Essential oil extraction, chemical analysis and anti-candida activity of calamintha nepeta (L.) Savi subsp. glandulosa (Req.) ball-new approaches. *Molecules* (2017) doi:10.3390/molecules22020203.

36. Božović, M., Navarra, A., Garzoli, S., Pepi, F. & Ragno, R. Esential oils extraction: a 24-hour steam distillation systematic methodology. *Nat. Prod. Res.* (2017) doi:10.1080/14786419.2017.1309534.

37. Garzoli, S. *et al.* Essential oil extraction, chemical analysis and anti-Candida activity of

Foeniculum vulgare Miller–new approaches. *Nat. Prod. Res.* (2018) doi:10.1080/14786419.2017.1340291.

38. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: From the natural environment to infectious diseases. *Nature Reviews Microbiology* vol. 2 95–108 (2004).

39. López, D., Vlamakis, H. & Kolter, R. Biofilms. *Cold Spring Harbor perspectives in biology* vol. 2 a000398 (2010).

40. Freires, I. A., Denny, C., Benso, B., De Alencar, S. M. & Rosalen, P. L. Antibacterial activity of essential oils and their isolated constituents against cariogenic bacteria: A systematic review. *Molecules* (2015) doi:10.3390/molecules20047329.

41. Manoharan, R. K., Lee, J. H., Kim, Y. G., Kim, S. Il & Lee, J. Inhibitory effects of the essential oils α-longipinene and linalool on biofilm formation and hyphal growth of Candida albicans. *Biofouling* (2017) doi:10.1080/08927014.2017.1280731.

42. Wang, Z. *et al.* Light controllable chitosan micelles with ROS generation and essential oil release for the treatment of bacterial biofilm. *Carbohydr. Polym.* (2019) doi:10.1016/j.carbpol.2018.10.095.

43. Artini, M. *et al.* Antimicrobial and antibiofilm activity and machine learning classification analysis of essential oils from different mediterranean plants against pseudomonas aeruginosa. *Molecules* **23**, (2018).

44. Patsilinakos, A. *et al.* Machine Learning Analyses on Data including Essential Oil Chemical Composition and In Vitro Experimental Antibiofilm Activities against Staphylococcus Species. *Molecules* (2019) doi:10.3390/molecules24050890.

45. Ragno, R. *et al.* Essential oils against bacterial isolates from cystic fibrosis patients by means of antimicrobial and unsupervised machine learning approaches. *Sci. Rep.* **10**, 1–11 (2020).

46. Kovach, K. *et al.* Evolutionary adaptations of biofilms infecting cystic fibrosis lungs promote mechanical toughness by adjusting polysaccharide production. *npj Biofilms Microbiomes* **3**, 1–9 (2017).

47. Alissa, E. M. & Ferns, G. A. Dietary fruits and vegetables and cardiovascular diseases risk. *Critical Reviews in Food Science and Nutrition* vol. 57 1950–1962 (2017).

48. Amagase, H. & Farnsworth, N. R. A review of botanical characteristics, phytochemistry, clinical relevance in efficacy and safety of Lycium barbarum fruit (Goji). *Food Research International* vol. 44 1702–1717 (2011).

49. Kafkaletou, M. *et al.* Nutritional value and consumer-perceived quality of fresh goji berries (Lycium barbarum L. and L. chinense L.) from plants cultivated in Southern Europe. *Fruits* (2018) doi:10.17660/th2018/73.1.1.

50. Potterat, O. Goji (Lycium barbarum and L. chinense): Phytochemistry, pharmacology and safety in the perspective of traditional uses and recent popularity. *Planta Medica* (2010) doi:10.1055/s-0029-1186218.

51. Jing Z. Dong. Analysis on the main active components of Lycium barbarum fruits and related environmental factors. *J. Med. Plants Res.* (2012) doi:10.5897/jmpr10.780.

52. Cheng, J. *et al.* An evidence-based update on the pharmacological activities and possible molecular targets of Lycium barbarum polysaccharides. *Drug design, development and therapy* (2015) doi:10.2147/DDDT.S72892.

53. Jin, M., Huang, Q., Zhao, K. & Shang, P. Biological activities and potential health benefit effects of polysaccharides isolated from Lycium barbarum L. *International Journal of Biological Macromolecules* (2013) doi:10.1016/j.ijbiomac.2012.11.023.

54. Mocan, A. *et al.* UHPLC-QTOF-MS analysis of bioactive constituents from two Romanian Goji (Lycium barbarum L.) berries cultivars and their antioxidant, enzyme inhibitory, and real-time cytotoxicological evaluation. *Food Chem. Toxicol.* (2018) doi:10.1016/j.fct.2018.01.054.

55. Wang, C. C., Chang, S. C., Inbaraj, B. S. & Chen, B. H. Isolation of carotenoids, flavonoids and polysaccharides from Lycium barbarum L. and evaluation of antioxidant activity. *Food Chem.* (2010) doi:10.1016/j.foodchem.2009.10.005.

56. Masci, A. *et al.* Lycium barbarum polysaccharides: Extraction, purification, structural characterisation and evidence about hypoglycaemic and hypolipidaemic effects. A review. *Food Chemistry* (2018) doi:10.1016/j.foodchem.2018.01.176.

57. Donno, D., Beccaro, G. L., Mellano, M. G., Cerutti, A. K. & Bounous, G. Goji berry fruit (Lycium spp.): Antioxidant compound fingerprint and bioactivity evaluation. *J. Funct.*

Foods (2015) doi:10.1016/j.jff.2014.05.020.

58. Zhang, Q., Chen, W., Zhao, J. & Xi, W. Functional constituents and antioxidant activities of eight Chinese native goji genotypes. *Food Chem.* (2016) doi:10.1016/j.foodchem.2016.01.046.

59. Zhou, Z. Q. *et al.* Polyphenols from wolfberry and their bioactivities. *Food Chem.* (2017) doi:10.1016/j.foodchem.2016.07.105.

60. Weller, P. & Breithaupt, D. E. Identification and Quantification of Zeaxanthin Esters in Plants using Liquid Chromatography-Mass Spectrometry. *J. Agric. Food Chem.* (2003) doi:10.1021/jf034803s.

61. Sajilata, M. G., Singhal, R. S. & Kamat, M. Y. The carotenoid pigment zeaxanthin - A review. in *Comprehensive Reviews in Food Science and Food Safety* (2008). doi:10.1111/j.1541-4337.2007.00028.x.

62. Cascella, R. *et al.* Age-Related Macular Degeneration: Insights into Inflammatory Genes. *Journal of Ophthalmology* (2014) doi:10.1155/2014/582842.

63. Desmarchelier, C. & Borel, P. Overview of carotenoid bioavailability determinants: From dietary factors to host genetic variations. *Trends in Food Science and Technology* (2017) doi:10.1016/j.tifs.2017.03.002.

64. Karioti, A., Bergonzi, M. C., Vincieri, F. F. & Bilia, A. R. Validated method for the analysis of Goji berry, a rich source of Zeaxanthin dipalmitate. *J. Agric. Food Chem.* (2014) doi:10.1021/jf503769s.

65. Hempel, J. *et al.* Ultrastructural deposition forms and bioaccessibility of carotenoids and carotenoid esters from goji berries (Lycium barbarum L.). *Food Chem.* (2017) doi:10.1016/j.foodchem.2016.09.065.

66. Clydesdale, F. M. & Ahmed, E. M. Colorimetry - methodology and applications. *C R C Crit. Rev. Food Sci. Nutr.* (1978) doi:10.1080/10408397809527252.

67. Pérez, F. & Granger, B. E. IPython: A System for Interactive Scientific Computing Python: An Open and General- Purpose Environment. *Comput. Sci. Eng.* (2007) doi:doi:10.1109/MCSE.2007.53.

68. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: A structure for

efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).

69. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* (2007) doi:10.1109/MCSE.2007.55.

70. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* (2010).

71. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Perform. Sci. Comput.* (2011).

72. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011) doi:10.1007/s13398-014-0173-7.2.

73. MA, W. ping, NI, Z. jing, LI, H. & CHEN, M. Changes of the Main Carotenoid Pigment Contents During the Drying Processes of the Different Harvest Stage Fruits of Lycium barbarum L. *Agric. Sci. China* (2008) doi:10.1016/S1671-2927(08)60077-2.

74. Rodriguez-Amaya, D. B. *A Guide to Carotenoid Analysis in Foods*. *Life Sciences* (2001).

75. Zheng, X. *et al.* A rapid and effective approach for on-site assessment of total carotenoid content in wolfberry juice during processing. *J. Sci. Food Agric.* (2015) doi:10.1002/jsfa.7038.

76. Peng, Y. *et al.* Quantification of zeaxanthin dipalmitate and total carotenoids in Lycium fruits (Fructus Lycii). *Plant Foods Hum. Nutr.* (2005) doi:10.1007/s11130-005-9550-5.

77. Turcsi, E., Nagy, V. & Deli, J. Study on the elution order of carotenoids on endcapped C18 and C30 reverse silica stationary phases. A review of the database. *Journal of Food Composition and Analysis* (2016) doi:10.1016/j.jfca.2016.01.005.

78. Fratianni, A. *et al.* Effect of a physical pre-treatment and drying on carotenoids of goji berries (Lycium barbarum L.). *LWT - Food Sci. Technol.* (2018) doi:10.1016/j.lwt.2018.02.048.

79. Humphries, J. M., Graham, R. D. & Mares, D. J. Application of reflectance colour measurement to the estimation of carotene and lutein content in wheat and triticale. *J. Cereal Sci.* (2004) doi:10.1016/j.jcs.2004.07.005.

80. Meléndez-Martínez, A. J., Britton, G., Vicario, I. M. & Heredia, F. J. Relationship

between the colour and the chemical structure of carotenoid pigments. *Food Chem.* (2007) doi:10.1016/j.foodchem.2006.03.015.

81. Cesa, S. *et al.* Evaluation of processing effects on anthocyanin content and colour modifications of blueberry (Vaccinium spp.) extracts: Comparison between HPLC-DAD and CIELAB analyses. *Food Chem.* (2017) doi:10.1016/j.foodchem.2017.03.153.

82. Kljak, K., Grbeša, D. & Karolyi, D. Reflectance colorimetry as a simple method for estimating carotenoid content in maize grain. *J. Cereal Sci.* (2014) doi:10.1016/j.jcs.2013.12.004.

83. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science (80-.).* (1983) doi:10.1126/science.220.4598.671.

84. Owen, A. B. A robust hybrid of lasso and ridge regression. in (2007). doi:10.1090/conm/443/08555.

85. Ramírez-Estrada, S., Borgatta, B. & Rello, J. Pseudomonas aeruginosa ventilator-associated pneumonia management. *Infection and Drug Resistance* (2016) doi:10.2147/IDR.S50669.

86. Kumar, A., Alam, A., Rani, M., Ehtesham, N. Z. & Hasnain, S. E. Biofilms: Survival and defense strategy for pathogens. *International Journal of Medical Microbiology* (2017) doi:10.1016/j.ijmm.2017.09.016.

87. Cag, Y., Caskurlu, H., Fan, Y., Cao, B. & Vahaboglu, H. Resistance mechanisms. *Annals of Translational Medicine* (2016) doi:10.21037/atm.2016.09.14.

88. Blanc, D. S., Francioli, P. & Zanetti, G. Molecular Epidemiology of Pseudomonas aeruginosa in the Intensive Care Units - A Review. *Open Microbiol. J.* **1**, 8–11 (2007).

89. Bassetti, M., Righi, E. & Viscoli, C. Pseudomonas aeruginosa Serious Infections: Mono or Combination Antimicrobial Therapy? *Curr. Med. Chem.* (2008) doi:10.2174/092986708783503186.

90. Mulcahy, L. R., Isabella, V. M. & Lewis, K. Pseudomonas aeruginosa Biofilms in Disease. *Microb. Ecol.* (2014) doi:10.1007/s00248-013-0297-x.

91. Bjarnsholt, T. *et al.* Why chronic wounds will not heal: A novel hypothesis. *Wound Repair and Regeneration* (2008) doi:10.1111/j.1524-475X.2007.00283.x.

92. Pradeep Kumar, S. S., Easwer, H. V. & Maya Nandkumar, A. Multiple Drug Resistant Bacterial Biofilms on Implanted Catheters - A Reservoir of Infection. *J. Assoc. Physicians India* (2013).

93. Papa, R. *et al.* Anti-biofilm activities from marine cold adapted bacteria against staphylococci and Pseudomonas aeruginosa. *Front. Microbiol.* (2015) doi:10.3389/fmicb.2015.01333.

94. Božovic, M., Pirolli, A. & Ragno, R. Mentha suaveolens Ehrh. (Lamiaceae) essential oil and its main constituent piperitenone oxide: Biological activities and chemistry. *Molecules* (2015) doi:10.3390/molecules20058605.

95. Božović, M., Ragno, R. & Tzakou, O. Calamintha nepeta (L.) Savi and its main essential oil constituent pulegone: Biological activities and chemistry. *Molecules* (2017) doi:10.3390/molecules22020290.

96. Langeveld, W. T., Veldhuizen, E. J. A. & Burt, S. A. Synergy between essential oil components and antibiotics: A review. *Critical Reviews in Microbiology* (2014) doi:10.3109/1040841X.2013.763219.

97. Stringaro, A. *et al.* Effects of mentha suaveolens essential oil alone or in combination with other drugs in candida albicans. *Evidence-based Complement. Altern. Med.* (2014) doi:10.1155/2014/125904.

98. Ganesh, P. S. & Rai, R. V. Inhibition of quorum-sensing-controlled virulence factors of Pseudomonas aeruginosa by Murraya koenigii essential oil: A study in a Caenorhabditis elegans infectious model. *J. Med. Microbiol.* (2016) doi:10.1099/jmm.0.000385.

99. Stojanović-Radić, Z., Pejčić, M., Stojanović, N., Sharifi-Rad, J. & Stanković, N. Potential of Ocimum basilicum L. and Salvia officinalis L. essential oils against biofilms of P. aeruginosa clinical isolates. *Cell. Mol. Biol.* (2016) doi:10.14715/cmb/2016.62.9.5.

100. Garzoli, S. *et al.* Multidisciplinary approach to determine the optimal time and period for extracting the essential oil from mentha suaveolens ehrh. *Molecules* (2015) doi:10.3390/molecules20069640.

101. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. F. & Nielsen, H. *Assessing the accuracy of prediction algorithms for classification: An overview*. *Bioinformatics* vol. 16

412–424 (2000).

102. Vert, J. P. Kernel methods in genomics and computational biology. in *Medical Informatics: Concepts, Methodologies, Tools, and Applications* (2008). doi:10.4018/978-1-60566-050-9.ch024.

103. Balasubramanian, D., Schneper, L., Kumari, H. & Mathee, K. A dynamic and intricate regulatory network determines Pseudomonas aeruginosa virulence. *Nucleic Acids Research* (2013) doi:10.1093/nar/gks1039.

104. Donlan, R. M. & Costerton, J. W. Biofilms: Survival mechanisms of clinically relevant microorganisms. *Clinical Microbiology Reviews* (2002) doi:10.1128/CMR.15.2.167-193.2002.

105. da Costa Krewer, C. *et al.* Resistance to antimicrobials and biofilm formation in Staphylococcus spp. isolated from bovine mastitis in the Northeast of Brazil. *Trop. Anim. Health Prod.* (2015) doi:10.1007/s11250-014-0752-9.

106. Furuya, E. Y. & Lowy, F. D. Antimicrobial-resistant bacteria in the community setting. *Nature Reviews Microbiology* (2006) doi:10.1038/nrmicro1325.

107. Satpathy, S., Sen, S. K., Pattanaik, S. & Raut, S. Review on bacterial biofilm: An universal cause of contamination. *Biocatalysis and Agricultural Biotechnology* (2016) doi:10.1016/j.bcab.2016.05.002.

108. Koo, H., Allan, R. N., Howlin, R. P., Stoodley, P. & Hall-Stoodley, L. Targeting microbial biofilms: Current and prospective therapeutic strategies. *Nature Reviews Microbiology* (2017) doi:10.1038/nrmicro.2017.99.

109. Otto, M. Staphylococcal biofilms. *Current Topics in Microbiology and Immunology* (2008) doi:10.1007/978-3-540-75418-3_10.

110. Li, Z. A Review of &lt;i&gt;Staphylococcus aureus&lt;/i&gt; and the Emergence of Drug-Resistant Problem. *Adv. Microbiol.* (2018) doi:10.4236/aim.2018.81006.

111. Nicholson, T. L., Shore, S. M., Smith, T. C. & Fraena, T. S. Livestock-Associated Methicillin-Resistant Staphylococcus aureus (LA-MRSA) Isolates of Swine Origin Form Robust Biofilms. *PLoS One* **8**, e73376 (2013).

112. Dohar, J. E. *et al.* Mucosal biofilm formation on middle-ear mucosa in a nonhuman

primate model of chronic suppurative otitis media. *Laryngoscope* (2005) doi:10.1097/01.mlg.0000172036.82897.d4.

113. Rogers, K. L., Fey, P. D. & Rupp, M. E. Coagulase-Negative Staphylococcal Infections. *Infectious Disease Clinics of North America* (2009) doi:10.1016/j.idc.2008.10.001.

114. Zhao, X. *et al.* Phenotype and RNA-seq-Based transcriptome profiling of Staphylococcus aureus biofilms in response to tea tree oil. *Microb. Pathog.* (2018) doi:10.1016/j.micpath.2018.07.027.

115. Vázquez-Sánchez, D., Galvão, J. A., Mazine, M. R., Gloria, E. M. & Oetterer, M. Control of Staphylococcus aureus biofilms by the application of single and combined treatments based in plant essential oils. *Int. J. Food Microbiol.* (2018) doi:10.1016/j.ijfoodmicro.2018.08.007.

116. Vaillancourt, K., LeBel, G., Yi, L. & Grenier, D. In vitro antibacterial activity of plant essential oils against Staphylococcus hyicus and Staphylococcus aureus, the causative agents of exudative epidermitis in pigs. *Arch. Microbiol.* (2018) doi:10.1007/s00203-018-1512-4.

117. Scaffaro, R., Lopresti, F., D'Arrigo, M., Marino, A. & Nostro, A. Efficacy of poly(lactic acid)/carvacrol electrospun membranes against Staphylococcus aureus and Candida albicans in single and mixed cultures. *Appl. Microbiol. Biotechnol.* (2018) doi:10.1007/s00253-018-8879-7.

118. Merghni, A. *et al.* Assessment of the antibiofilm and antiquorum sensing activities of Eucalyptus globulus essential oil and its main component 1,8-cineole against methicillin-resistant Staphylococcus aureus strains. *Microb. Pathog.* (2018) doi:10.1016/j.micpath.2018.03.006.

119. Kot, B., Wierzchowska, K., Grużewska, A. & Lohinau, D. The effects of selected phytochemicals on biofilm formed by five methicillin-resistant Staphylococcus aureus. *Nat. Prod. Res.* (2018) doi:10.1080/14786419.2017.1340282.

120. Nostro, A. *et al.* Effects of oregano, carvacrol and thymol on Staphylococcus aureus and Staphylococcus epidermidis biofilms. *J. Med. Microbiol.* (2007) doi:10.1099/jmm.0.46804-0.

121. Chovanová, R., Mikulášová, M. & Vaverková, Š. In vitro antibacterial and antibiotic resistance modifying effect of bioactive plant extracts on methicillin-resistant staphylococcus epidermidis. *Int. J. Microbiol.* (2013) doi:10.1155/2013/760969.

122. Chovanová, R., Mezovská, J., Vaverková & Mikulášová, M. The inhibition the Tet(K) efflux pump of tetracycline resistant Staphylococcus epidermidis by essential oils from three Salvia species. *Lett. Appl. Microbiol.* (2015) doi:10.1111/lam.12424.

123. Riahi, S., Pourbasheer, E., Ganjali, M. R. & Norouzi, P. Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: Concerns to support vector machine. *J. Hazard. Mater.* (2009) doi:10.1016/j.jhazmat.2008.11.097.

124. Taghadomi-Saberi, S., Garcia, S. M., Masoumi, A. A., Sadeghi, M. & Marco, S. Classification of bitter orange essential oils according to fruit ripening stage by untargeted chemical profiling and machine learning. *Sensors (Switzerland)* (2018) doi:10.3390/s18061922.

125. Drevinskas, T. *et al.* Confirmation of the antiviral properties of medicinal plants via chemical analysis, machine learning methods and antiviral tests: A methodological approach. *Anal. Methods* (2018) doi:10.1039/c8ay00318a.

126. Heilmann, C., Gerke, C., Perdreau-Remington, F. & Götz, F. Characterization of Tn917 insertion mutants of Staphylococcus epidermidis affected in biofilm formation. *Infect. Immun.* (1996) doi:10.1128/iai.64.1.277-282.1996.

127. Humphries, R. M. *et al.* CLSI methods development and standardization working group best practices for evaluation of antimicrobial susceptibility tests. *Journal of Clinical Microbiology* (2018) doi:10.1128/JCM.01934-17.

128. Perkel, J. M. Programming: Pick up Python. *Nature* (2015) doi:10.1038/518125a.

129. Kluyver, T. *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. *Position. Power Acad. Publ. Play. Agents Agendas* 87–90 (2016) doi:10.3233/978-1-61499-649-1-87.

130. Wei, P., Lu, Z. & Song, J. Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety* (2015) doi:10.1016/j.ress.2015.05.018.

131. Kramer, A. *et al.* datascienceinc/Skater: 1.1.2. (2018) doi:10.5281/ZENODO.1423046.

132. Klopman, G. & Kalos, A. N. Causality in structure—activity studies. *J. Comput. Chem.* (1985) doi:10.1002/jcc.540060520.

133. Sandasi, M., Leonard, C. M. & Viljoen, A. M. The effect of five common essential oil components on Listeria monocytogenes biofilms. *Food Control* (2008) doi:10.1016/j.foodcont.2007.11.006.

134. Sandasi, M., Leonard, C. M. & Viljoen, A. M. The in vitro antibiofilm activity of selected culinary herbs and medicinal plants against Listeria monocytogenes. *Lett. Appl. Microbiol.* **50**, 30–35 (2010).

135. Negreiros, M. de O. *et al.* Antimicrobial and antibiofilm activity of Baccharis psiadioides essential oil against antibiotic-resistant Enterococcus faecalis strains. *Pharm. Biol.* (2016) doi:10.1080/13880209.2016.1223700.

136. Szczepanski, S. & Lipski, A. Essential oils show specific inhibiting effects on bacterial biofilm formation. *Food Control* (2013) doi:10.1016/j.foodcont.2013.08.023.

137. Kannappan, A. *et al.* Inhibitory efficacy of geraniol on biofilm formation and development of adaptive resistance in Staphylococcus epidermidis RP62A. *J. Med. Microbiol.* (2017) doi:10.1099/jmm.0.000570.

138. Nuryastuti, T. *et al.* Effect of cinnamon oil on icaA expression and biofilm formation by Staphylococcus epidermidis. *Appl. Environ. Microbiol.* (2009) doi:10.1128/AEM.00875-09.

139. Chueca, B., Pagán, R. & García-Gonzalo, D. Differential mechanism of Escherichia coli inactivation by (+)-limonene as a function of cell physiological state and drug's concentration. *PLoS One* (2014) doi:10.1371/journal.pone.0094072.

140. Rummaneethorn, N., Caoili, C. M., Yeung-Cheung, A. K. & Pappas, C. J. D-limonene Increases Efficacy of Rifampicin as an Inhibitor of In Vitro Growth of Opportunistic Staphylococcus epidermidis RP62A. in *The National Conference On undergraduate Research (NCUR) 2016* (2016).

141. Subramenium, G. A., Vijayakumar, K. & Pandian, S. K. Limonene inhibits streptococcal biofilm formation by targeting surface-associated virulence factors. *J. Med. Microbiol.*

(2015) doi:10.1099/jmm.0.000105.

142. Cerioli, M. F., Moliva, M. V., Cariddi, L. N. & Reinoso, E. B. Effect of the essential oil of Minthostachys verticillata (Griseb.) epling and limonene on biofilm production in pathogens causing bovine mastitis. *Front. Vet. Sci.* (2018) doi:10.3389/fvets.2018.00146.

143. Espina, L., Pagán, R., López, D. & García-Gonzalo, D. Individual constituents from essential oils inhibit biofilm mass production by multi-drug resistant staphylococcus aureus. *Molecules* (2015) doi:10.3390/molecules200611357.

144. Majumdar, S. & Mondal, S. Perspectives on Quorum sensing in Fungi. *bioRxiv* (2015) doi:10.1101/019034.

145. Cafiso, V. *et al.* agr-Genotyping and transcriptional analysis of biofilm-producing Staphylococcus aureus. *FEMS Immunol. Med. Microbiol.* (2007) doi:10.1111/j.1574-695X.2007.00298.x.

146. Vuong, C., Gerke, C., Somerville, G. A., Fischer, E. R. & Otto, M. Quorum-Sensing Control of Biofilm Factors in Staphylococcus epidermidis . *J. Infect. Dis.* (2003) doi:10.1086/377239.

147. Büttner, H., Mack, D. & Rohde, H. Structural basis of Staphylococcus epidermidis biofilm formation: Mechanisms and molecular interactions. *Frontiers in Cellular and Infection Microbiology* (2015) doi:10.3389/fcimb.2015.00014.

148. Harris, A. & Argent, B. E. The cystic fibrosis gene and its product CFTR. *Semin. Cell Dev. Biol.* (1993).

149. Anderson, G. G. Pseudomonas aeruginosa biofilm formation in the CF lung and its implications for therapy. in *Cystic Fibrosis-Renewed Hopes Through Research, ed Sriramulu D., editor.(London: InTech* 153–180 (2012). doi:10.5772/30529.

150. Gibson, R. L., Burns, J. L. & Ramsey, B. W. Pathophysiology and Management of Pulmonary Infections in Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine* (2003) doi:10.1164/rccm.200304-505SO.

151. Hauser, A. R., Jain, M., Bar-Meir, M. & McColley, S. A. Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev.* (2011)

doi:10.1128/CMR.00036-10.

152. Malhotra, S., Limoli, D. H., English, A. E., Parsek, M. R. & Wozniak, D. J. Mixed communities of mucoid and nonmucoid Pseudomonas aeruginosa exhibit enhanced resistance to host antimicrobials. *MBio* (2018) doi:10.1128/mBio.00275-18.

153. MacKenzie, T. *et al.* Longevity of patients with cystic fibrosis in 2000 to 2010 and beyond: Survival analysis of the Cystic Fibrosis Foundation Patient Registry. *Ann. Intern. Med.* (2014) doi:10.7326/M13-0636.

154. Molchanova, N., Hansen, P. R. & Franzyk, H. Advances in development of antimicrobial peptidomimetics as potential drugs. *Molecules* (2017) doi:10.3390/molecules22091430.

155. Smith, W. D. *et al.* Current and future therapies for Pseudomonas aeruginosa infection in patients with cystic fibrosis. *FEMS Microbiology Letters* (2017) doi:10.1093/femsle/fnx121.

156. Perez, L. R. R. & Barth, A. L. Biofilm production using distinct media and antimicrobial susceptibility profile of Pseudomonas aeruginosa. *Brazilian J. Infect. Dis.* (2011) doi:10.1016/S1413-8670(11)70196-9.

157. Mikkelsen, H., McMullan, R. & Filloux, A. The Pseudomonas aeruginosa reference strain PA14 displays increased virulence due to a mutation in ladS. *PLoS One* (2011) doi:10.1371/journal.pone.0029113.

158. Kerem, E. *et al.* Standards of care for patients with cystic fibrosis: A European consensus. *J. Cyst. Fibros.* (2005) doi:10.1016/j.jcf.2004.12.002.

159. Levin, T. P., Suh, B., Axelrod, P., Truant, A. L. & Fekete, T. Potential clindamycin resistance in clindamycin-susceptible, erythromycin-resistant Staphylococcus aureus: Report of a clinical failure. *Antimicrob. Agents Chemother.* (2005) doi:10.1128/AAC.49.3.1222-1224.2005.

160. Palzkill, T. Metallo-β-lactamase structure and function. *Ann. N. Y. Acad. Sci.* (2013) doi:10.1111/j.1749-6632.2012.06796.x.

161. Meletis, G. & Bagkeri, M. Pseudomonas aeruginosa: Multi-Drug-Resistance Development and Treatment Options. in *Infection Control* (2013). doi:10.5772/55616.

162. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* (1987) doi:10.1016/0377-0427(87)90125-7.

163. López-Causapé, C., Rojo-Molinero, E., MacIà, M. D. & Oliver, A. The problems of antibiotic resistance in cystic fibrosis and solutions. *Expert Review of Respiratory Medicine* (2015) doi:10.1586/17476348.2015.995640.

164. Murray, J. L., Kwon, T., Marcotte, E. M. & Whiteley, M. Intrinsic antimicrobial resistance determinants in the superbug pseudomonas aeruginosa. *MBio* (2015) doi:10.1128/mBio.01603-15.

165. Dodémont, M. *et al.* Emergence of livestock-associated MRSA isolated from cystic fibrosis patients: Result of a Belgian national survey. *J. Cyst. Fibros.* (2019) doi:10.1016/j.jcf.2018.04.008.

166. Cegelski, L., Marshall, G. R., Eldridge, G. R. & Hultgren, S. J. The biology and future prospects of antivirulence therapies. *Nature Reviews Microbiology* (2008) doi:10.1038/nrmicro1818.

167. Topa, S. H. *et al.* Cinnamaldehyde disrupts biofilm formation and swarming motility of pseudomonas aeruginosa. *Microbiol. (United Kingdom)* (2018) doi:10.1099/mic.0.000692.

168. Vasireddy, L., Bingle, L. E. H. & Davies, M. S. Antimicrobial activity of essential oils against multidrug-resistant clinical isolates of the burkholderia cepacia complex. *PLoS One* (2018) doi:10.1371/journal.pone.0201835.

169. Poma, P. *et al.* Essential oil composition of alluaudia procera and in vitro biological activity on two drug-resistant models. *Molecules* (2019) doi:10.3390/molecules24162871.

170. Karumathil, D. P., Nair, M. S., Gaffney, J., Kollanoor-Johny, A. & Venkitanarayanan, K. Trans-Cinnamaldehyde and eugenol increase Acinetobacter baumannii sensitivity to beta-lactam antibiotics. *Front. Microbiol.* (2018) doi:10.3389/fmicb.2018.01011.

171. Rosato, A. *et al.* Elucidation of the synergistic action of Mentha Piperita essential oil with common antimicrobials. *PLoS One* (2018) doi:10.1371/journal.pone.0200902.

172. Tetard, A., Zedet, A., Girard, C., Plésiat, P. & Llanes, C. Cinnamaldehyde induces

expression of efflux pumps and multidrug resistance in pseudomonas aeruginosa. *Antimicrob. Agents Chemother.* (2019) doi:10.1128/AAC.01081-19.