# Classification of common terpenoids based on Aromaticity and Molecular Size

Diego Soeldi, 2025-12-03

## Introduction

In previous works it was stated that aromaticity and molecular size are good indicators to differentiate between the different types of metabolites in a metabolome [1]. This was concluded from multiple different descriptors being combined in a principal component analysis (PCA), based on Aromaticity, Size and Lipophilicity. This work showed that in the class of terpenoids, the heterogeneity in aromaticity and size is substantial which is utilized to train a Machine Learning model that differentiates between the 5 most common terpenoids found within the metabolome of tropical plants. These terpenoids were chosen because they were the 5 biggest groups of metabolites found in the given data [2].

## Methods

The initial data was a part of the work by Walker et al. and described different molecules found in tropical plants [1,2]. After initial data exploration, multiple descriptors, chosen for their relation to aromaticity and molecule size and calculated with the *rdkit* library for python, were chosen for the training and analysis process. A *Random Forest* (RF) model was trained using *randomized search CV* to optimize the hyperparameters. The precision score, calculated from predicted terpenoid lables, using separate test data, is shown in the left bar plot in figure 1. Additionally, a permutation importance method was used to find the importance of each feature based on its impact when it gets left out. This was also plotted as a bar plot; visible on the right in Figure 1. The Scripts for this analysis can be found on Github [3].

## Results

Precision values of 0.75 to 0.96 were produced in all categories. The permutation analysis shows that the molecular weight was the most important feature used in the classification, but all features added to the analysis, and nothing could be left out without impacting the final accuracy.

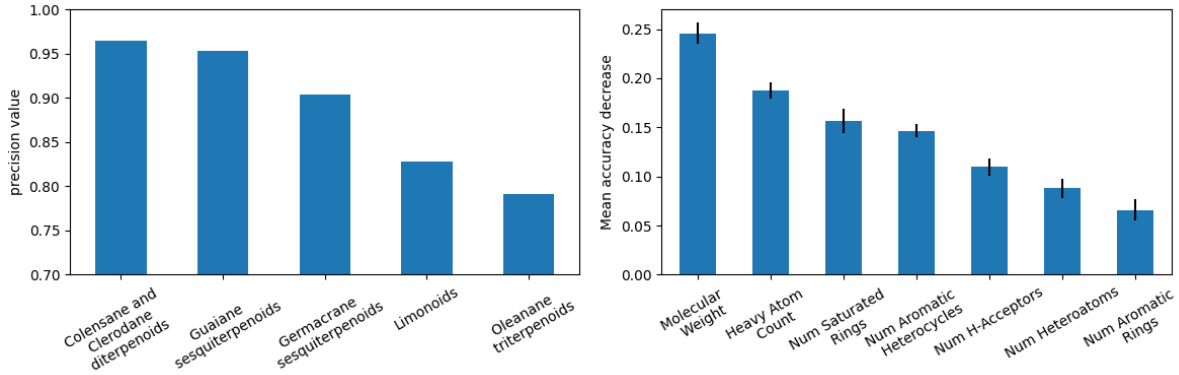


*Figure 1 Precision values of the test data (left) and feature importance of used training features (right) shown as separate box plots. Precision value was calculated using a separate test set to test the models performance on. The feature importance analysis was calculated using a Permutation analysis method and contains standard deviations for each analyzed missing feature.*

## Discussion

The importance of each descriptor shows that descriptors that relate to aromaticity and size of the molecule both impact the classification of terpenoids. This is in line with the findings of Walker et al. as the PCA showed similar findings where different classes of metabolites could be differentiated from each other [1]. This subset of Features describes Colensane and Clerodane diterpenoids best; the reason for this is hard to make out, as interpreting *RF* classifiers can get complicated, due to its large number of decision trees and the presence of multiple predictors in the data set [4]. In the end it can be said that molecular size and aromaticity can be used to differentiate the most common terpenoid types in the given data. This can be expanded upon to differentiate metabolites from each other.

## References

[1] T. W. N. Walker *et al.*, "Leaf metabolic traits reveal hidden dimensions of plant form and function," *SCIENCE ADVANCES*, doi: 10.1126/sciadv.adi4029.
[2] T. Walker, „Data from: Leaf metabolic traits reveal hidden dimensions of plant form and function". Zenodo, Juli 31, 2023. doi: 10.5281/zenodo.8160741.
[3] Diego Söldi, DSA_103_projects_dsoeldi, 2025, https://github.com/DSoeldi/DSA_103_projects_dsoeldi
[4] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Machine Learning with Applications*, vol. 6, p. 100094, Dec. 2021, doi: 10.1016/j.mlwa.2021.100094.