Machine Learning Analysis Reveals Correlation between Lipophilicity and Combination of Molecular Traits.

Hugo Peeters, 2025-12-02

Introduction

The metabolites in plants are a complex system of molecules which are involved in the biochemical processes in plants. These metabolites have given information on many topics including the plant's defense mechanisms, nutritional values, crop yield, medicinal properties and more.[1–3] This makes studying the plant's metabolome of utmost importance.

The data used for the analysis in this paper comes from a previous study on metabolic features in plants, and was used to find and group the different correlations in the plants metabolome.[4] The questions asked in this paper were if there are any correlations in three different predicted molecular traits, lipophilicity, number of heteroatoms and molecular weight, and if we could use machine learning to support this correlation.

To answer these questions, we used a python script which plots the correlation plots of the three molecular traits (Figure 1). We then used random forest, which is a machine learning technique used for classification, regression and more,[5] to check if there is a correlation or not.

Methods

The approach taken for this analysis was to try and understand the R code used by the people who previously analyzed this data.[4] ChatGPT was used for this step as the complexity and difficulty of the R code was beyond our understanding of R. The code was then partially translated to python with the help of ChatGPT and adjusted it to align with the analysis done in this project. The source data and code accompanying the data analysis can be found in the GitHub repository https://github.com/HugoPeeters25/DSA103_project.

Results

The correlation between the three descriptors were visualized as shown in Figure 1. The correlation plot does not show much of a correlation between the lipophilicity and the other two molecular traits. To test if there was any correlation, machine learning in the form of random forest regression was used, which showed an R^2 of 0.87.

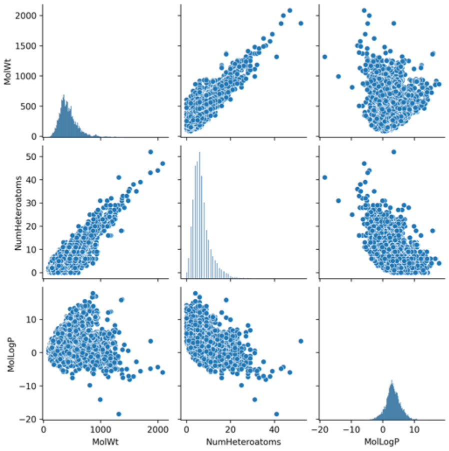

Figure 1 Pair plot showing the individual correlations between the three molecular traits.

Discussion

An R^2 value of 0.87 means that the random forest was able to roughly identify the lipophilicity by only looking at the molecular weight and number of heteroatoms. This means that there is a correlation between the lipophilicity of a molecule and the molecular weight combined with the number of hetero atoms, supporting our hypothesis. This adds to what the original paper[4] said, when they grouped these molecular traits into five groups, showing that not just the different molecular traits can be used for correlation, but also combinations of seemingly uncorrelated traits.

References

(1) Lu, J.; Xia, S. Metabolomics and Plant Defense. *Metabolites* **2025**, *15* (3), 171. https://doi.org/10.3390/metabo15030171.

(2) Salam, U.; Ullah, S.; Tang, Z.-H.; Elateeq, A. A.; Khan, Y.; Khan, J.; Khan, A.; Ali, S. Plant Metabolomics: An Overview of the Role of Primary and Secondary Metabolites against Different Environmental Stress Factors. *Life Basel Switz.* **2023**, *13* (3), 706. https://doi.org/10.3390/life13030706.

(3) Marchev, A. S.; Vasileva, L. V.; Amirova, K. M.; Savova, M. S.; Balcheva-Sivenova, Z. P.; Georgiev, M. I. Metabolomics and Health: From Nutritional Crops and Plant-Based Pharmaceuticals to Profiling of Human Biofluids. *Cell. Mol. Life Sci. CMLS* **2021**, *78* (19–20), 6487–6503. https://doi.org/10.1007/s00018-021-03918-3.

(4) Walker, T. W. N.; Schrodt, F.; Allard, P.-M.; Defossez, E.; Jassey, V. E. J.; Schuman, M. C.; Alexander, J. M.; Baines, O.; Baldy, V.; Bardgett, R. D.; Capdevila, P.; Coley, P. D.; van Dam, N. M.; David, B.; Descombes, P.; Endara, M.-J.; Fernandez, C.; Forrister, D.; Gargallo-Garriga, A.; Glauser, G.; Marr, S.; Neumann, S.; Pellissier, L.; Peters, K.; Rasmann, S.; Roessner, U.; Salguero-Gómez, R.; Sardans, J.; Weckwerth, W.; Wolfender, J.-L.; Peñuelas, J. Leaf Metabolic Traits Reveal Hidden Dimensions of Plant Form and Function. *Sci. Adv.* **2023**, *9* (35), eadi4029. https://doi.org/10.1126/sciadv.adi4029.

(5) Random Forest. *Wikipedia*; 2025.