DSA103 Data Project Julian Patrick Stoerr HS25

01.12.2025

Introduction

Plant metabolites have been widely studied, and many of their functions have been linked to plant physiology and ecosystems. However, how plant metabolism varies systematically across species remains poorly characterised. Walker et al. (2023) addressed this by analysing leaf chemistry using chemical properties of metabolites, introducing a biochemical perspective on functional metabolic traits. In this project, tropical plant metabolite data were analysed using chemical descriptors extracted with RDKit from provided SMILES structures. The main question was whether unsupervised learning could identify meaningful molecular patterns from chemical properties alone, aiming to systematically classify or understand the underlying biochemical strategies implemented by plants.

Methods

Annotated metabolite features with SMILES strings were retained for analysis, and entries with missing or empty structural information were removed. Identical molecular structures were grouped by unique SMILES code to obtain one descriptor profile per compound. Chemical properties were extracted from the molecules, standardised, and Principal Component Analysis (PCA) was performed using the scikit-learn library to summarise global structure. To identify group structure, the KMeans clustering algorithm was applied in PCA space and the number of clusters was evaluated using silhouette analysis. Cluster characteristics were summarised by computing mean descriptor values per cluster, and relationships were explored using a correlation matrix. All analyses were performed in Python using pandas, scikit-learn, seaborn, matplotlib, and RDKit. The data preparation protocol used the provided “derive_chemistry.R” script from the original publication.

Results

The PCA revealed structured variation in chemical descriptor space, with the first principal component capturing approximately 18% and the second 10% of total variance. The distribution showed continuous overlap rather than distinct clusters, indicating that metabolites form a spectrum rather than discrete categories. Silhouette analysis supported a three-cluster solution as the most appropriate grouping, revealing three dominant chemotypes with distinct chemical profiles. These chemotypes were visualised using a bar-plot summarising mean descriptor values per cluster.

Discussion

The results demonstrated that the chemical properties of the analysed molecules themselves capture biologically meaningful organisation in tropical plant metabolites. The conducted analysis supports the findings of Walker et al. that plant metabolism reflects structured biochemical strategies rather than random molecular signals. The continuous distribution of the PCA indicates that chemical traits form gradients and not discrete categories. This suggests that metabolic strategies are flexible and might overlap in different kingdoms and species. Overall, this analysis shows that metabolic chemistry provides a complex and meaningful layer of functional information in plants, which could be a valuable tool for further research into patterns invisible in macroscopic analysis.

Sources

Babitz, Kevin. “Introduction to K-Means Clustering with Scikit-Learn in Python.” *Datacamp.com*, DataCamp, 5 July 2018, www.datacamp.com/tutorial/k-means-clustering-python.

GeeksforGeeks. “Implementing PCA in Python with Scikitlearn.” *GeeksforGeeks*, 16 Feb. 2021, www.geeksforgeeks.org/machine-learning/implementing-pca-in-python-with-scikit-learn/.

“RDKit.” *GitHub*, 28 Aug. 2022, github.com/rdkit/rdkit.

Walker, T K, et al. “Leaf Metabolic Traits Reveal Hidden Dimensions of Plant Form and Function.” *Science Advances*, vol. 9, no. 35, 1 Sept. 2023, https://doi.org/10.1126/sciadv.adi4029. Accessed 18 Sept. 2023.

Walker, Tom. “Data From: Leaf Metabolic Traits Reveal Hidden Dimensions of Plant Form and Function.” *Zenodo (CERN European Organization for Nuclear Research)*, 18 July 2023, https://doi.org/10.5061/dryad.zpc866tdn. Accessed 1 Dec. 2025.