Exploring Chemical Space in Tropical Plant Metabolites

Principal Component Analysis and Probability-Density Visualization

Maurice Cedric Blasko

05.12.2025

1 Introduction

Chemical diversity in plant metabolomes provides insight into ecological and physiological strategies. Walker et al. (2023) showed that metabolomic variation across ∼800 species can be summarized along a small number of chemical axes. This project evaluates whether simple descriptors derived solely from molecular formulas can recover comparable structure within the tropical subset of their dataset. Because conventional PCA scatter plots suffer from overplotting, kernel density estimation (KDE) was applied to improve visualization of class-level chemical space.

2 Methods

The file `mtbs_tropical_annotations.tsv` contained molecular formulas and NPClassifier superclasses for tropical plant metabolites. Formulas were parsed into atom counts for C, H, N, O, P, S, and Other", and the derived descriptors *Total_atoms* and *Hetero_atoms*; entries without valid formulas were removed. The eight most frequent NPClassifier superclasses were retained. All remaining categories were merged into Other". Descriptor values were standardized and subjected to PCA, where PC1 primarily reflected molecular size and PC2 heteroatom enrichment. Visualization included a scatter plot and a KDE-based probability-density representation to provide visual clarity. Figures were generated in Python and exported as PNGs.

3 Results

The PCA revealed two interpretable axes. PC1 represented variation in molecular size and carbon content, while PC2 represented heteroatom enrichment. Scatter plots showed partial separation among superclasses. The great number of datapoints obscured distributional structure though. The KDE visualization was done to provided improved visual clarity.

Terpenoid-related superclasses were concentrated at high PC1 values, reflecting large carbon-rich structures. Flavonoids and isoflavonoids shifted toward higher PC2, consistent with oxygen-rich substitution. Tryptophan alkaloids displayed displacement along PC2 consistent with nitrogen incorporation. The "Other" category was diffuse, as expected for a heterogeneous chemical grouping.

4 Discussion

Simple molecular-formula descriptors recovered chemically meaningful axes comparable to those described by Walker et al. (2023). KDE-based density visualization improved the interpretability of class-level structure, even though high data volume still limits the resolution of individual categories. Despite this limitation, probability-density approaches remain useful for summarizing complex metabolomic datasets, particularly when scatter-based representations obscure distributional patterns.

Figures

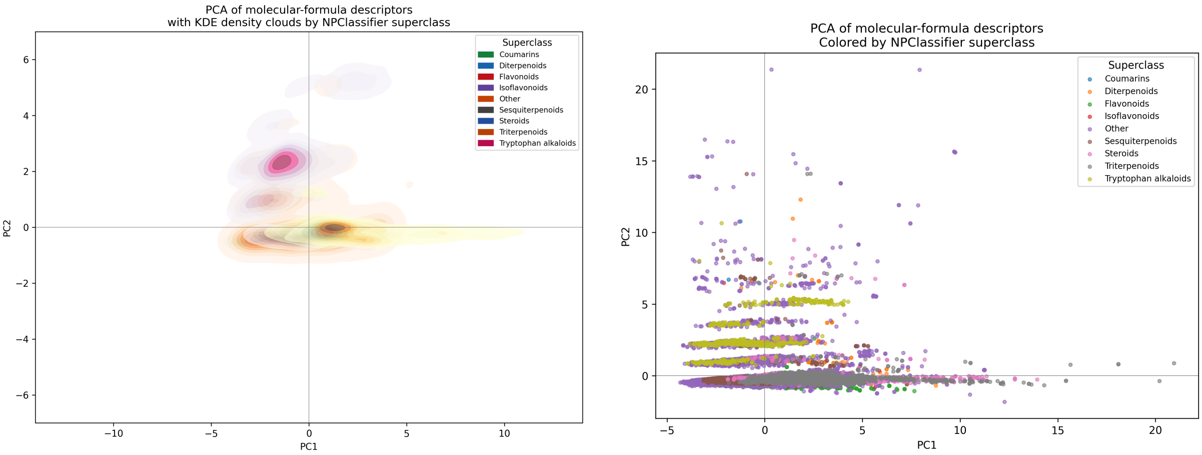

Figure 1: PCA of molecular-formula descriptors. **Left:** KDE-based probability-density clouds for major NPClassifier superclasses. **Right:** Scatter plot of individual metabolites colored by superclass. Both share the same PC1–PC2 coordinates.

References

Walker et al. (2023), *Science Advances*, "Leaf metabolic traits reveal hidden dimensions of plant form and function."
Example Data Story by Diego Söldi (2024).