Plants: Natural Products and Chemical Space

Hassan Haydar 02.12.2025

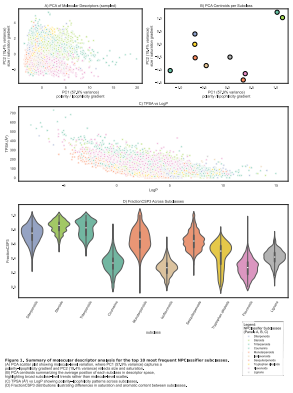
Introduction

Natural products are structurally diverse metabolites central to plant function and important in medicine, nutrition, and biotechnology. The metabolite data set from Walker et al. (2023) [1], originally used to explore how metabolic traits relate to plant form and function, captures a wide range of metabolite structural diversity. This analysis investigates whether major natural-product families (NPClassifier subclasses) show distinct chemical properties. I focus on three key descriptors: TPSA (polarity), LogP (lipophilicity), and FractionCSP3 (carbon saturation) calculated with RDKit. By comparing these descriptors and applying PCA, I examine how subclasses differ in polarity and saturation and whether they occupy distinct regions of chemical space.

Methods

The file *mtbs_tropical_annotations.tsv* from *Walker et al. (2023)* [1] was imported into Python and NPClassifier annotations together with the SMILES structures were retained. All entries lacking a valid SMILES were removed. Molecules sharing the same SMILES were identified and collapsed by grouping on the SMILES string and selecting the most common NPClassifier annotations ("class", "subclass", "my_class"). Column names were simplified for convenience, and each unique structure was assigned a structural ID (SID). Only the top 10 most frequent NPClassifier subclasses were included to maintain readability. For every unique SMILES, a set of basic molecular descriptors was calculated using RDKit: molecular weight, LogP, TPSA, hydrogen bond donors and acceptors, ring count, and fraction of sp^3 carbons. The resulting descriptor table (df_final) was saved as *mtbs_tropical_descriptors.csv* for subsequent analysis. PCA (scikit-learn) was used to summarize variation across these descriptors, and three descriptor-based visualizations were creared: PCA scatterplot, PCA subclass centroids, a TPSA vs. LogP plot, and a FractionCSP3 violin plot. All analyses were conducted in python (final_project.ipynb)

Results



Discussion

The results show that the top 10 NPClassifier subclasses differ in broad chemical traits. PCA indicates that most variation among metabolites follows two main axes polarity/lipophilicity and molecular size/saturation consistent with the coordinated chemical trait structure described in Fig. 2 of *Walker et. al.* [1]. The TPSA vs. LogP plot shows that subclasses vary in how polar or hydrophobic their metabolites tend to be, aligning with the established role of TPSA as a predictor of transport and absorption in medicinal chemistry [2]. The FractionCSP3 violin plot shows that there is a difference in carbon saturation across subclasses, higher saturation (higher Fsp^3) is often associated with improved solubility and better compound progression in drug discovery [3].

References

[1] Walker *et al*., *Sci. Adv.* **2023**, 9, eadi4029.

[2] *Lovering, Bikker, and Humblet, J. Med. Chem.* ***2009****, 52, 6752–6756*

[3] Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714–3717.