Data Story Template DSA103 project

Suvam Kumar Das, 2025-12-05

Introduction

Tropical natural products represent one of the richest reservoirs of unexplored chemical diversity, and the dataset from Walker TWN et al[1]. namely *mtbs_tropical_annotations.tsv*. I used a detailed cheminformatics workflow to answer "How is the chemical diversity of tropical natural products structured across biosynthetic classes, physicochemical properties, and molecular scaffolds[2]?" question. I started by calculating structural descriptors and fingerprints for each compound. Next, I used dimensionality-reduction methods like PCA and UMAP to find overall patterns in chemical space. I also looked at scaffold distributions and property trends within each class to spot key features of major natural-product families. This combined approach gives both broad and detailed insights into how tropical natural product diversity is organized.

Methods

The analysis began by loading and cleaning the dataset, removing invalid or duplicate SMILES. Using RDKit[3], molecular descriptors and Morgan fingerprints were computed for each compound. After scaling descriptors, PCA and hierarchical clustering were performed to reveal major property patterns. Murcko scaffolds[2] and class distributions were analyzed to summarize chemical families. UMAP embedding of fingerprints visualized structural similarity, with QED[4] values overlaid. All plots and processed data were saved, and the fully commented code and metadata ensure reproducibility.

Results

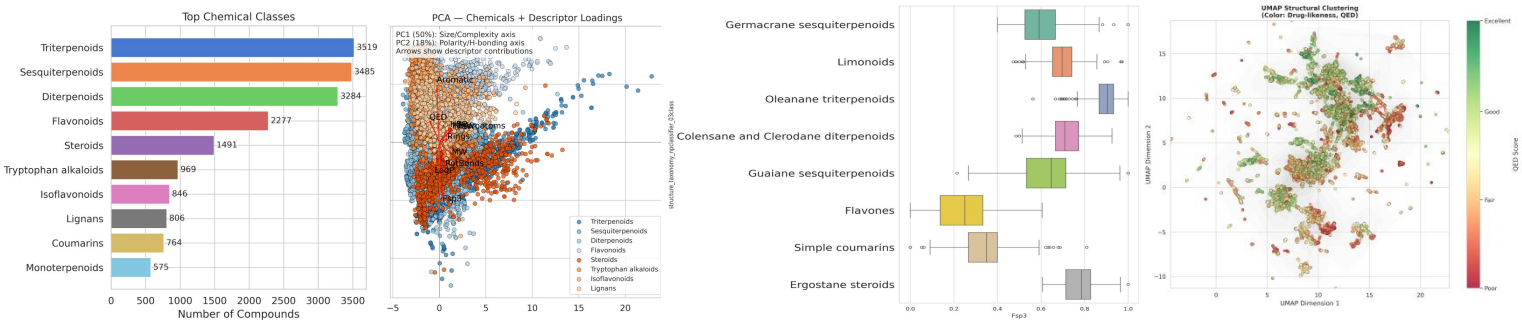

Figure 1. Summary of chemical diversity in the *mtbs_tropical_annotations.tsv* dataset.
The panels show class frequencies, descriptor-based PCA, property correlations for Fsp3, and a UMAP structural embedding colored by QED, together illustrating key trends in tropical natural-product space.

Discussion

The analysis reveals that tropical natural products in this dataset are structurally diverse, with variation largely driven by size, polarity, and rigidity, consistent with Walker et al[1]. Correlation and PCA analyses highlight known relationships among molecular properties, while scaffold and UMAP analyses show clustering of characteristic natural-product cores. Our results closely parallel the metabolic-trait patterns reported by Walker et al.[1], who showed that plant metabolomes organize along major chemical axes reflecting molecular size/complexity and polarity. Similarly, our PCA and UMAP analyses reveal that these descriptors drive the principal structural differentiation within the dataset. The predominance of terpenoids, together with their high Fsp^3 values, is consistent with established natural-product chemical-space trends, in which terpenoid scaffolds exhibit high three-dimensionality and structural complexity[5]. The UMAP projection shows clear structural clusters, with high-QED compounds grouping into defined regions, consistent with QED-driven organization of chemical space[6]. Overall, these compounds occupy a complex, biologically relevant chemical space, useful for novel scaffold discovery but often outside typical drug-like ranges.

References

1. T. W. N. Walker et al., Leaf metabolic traits reveal hidden dimensions of plant form and function. Sci. Adv. 9, eadi4029 (2023).
2. M. A. Murcko, The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 39, 2887–2893 (1996).
3. G. Landrum, RDKit: Open-source cheminformatics. Available at: https://www.rdkit.org (2023).
4. G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. Molecular frameworks. J. Med. Chem. 39, 2887–2893 (1996).
5. H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery. Drug Discov. Today 23, 1241–1250 (2018).
6. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, A. L. Hopkins, Quantifying the chemical beauty of drugs. Nat. Chem. 4, 90–98 (2012).