The outliers were output as a list of SMILES strings, which were converted to RDKit objects. Within the outliers dataset there was one invalid SMILES, which was excluded leaving 41 structures. A quick visual examination of the structures showed no apparent similarities between the structures, so a clustering analysis was completed.1

Using RDKit, RDK5 molecular fingerprints were created for each compound, each containing 2048 possible fragments. The similarity between two fingerprints was calculated using the Tanimoto Coefficient.2 Completing this for all pairs of fingerprints allowed creation of a similarity matrix between all molecular fingerprint pairs. Subtraction of each of these values from 1 created a distance matrix between all molecular fingerprint pairs.

Butina clustering was completed on the distance matrix using RDKit.3 Experimenting with varying cut-offs from 0.0 to 1.0 in steps of 0.2 found that a cut-off of at least 0.8 was required for clusters of at least five compounds. Visual examination of the two largest clusters showed all compounds contained aromatic systems. Further work is required to determine how to improve the model to account for aromatic compounds. The initial step will be to determine if the model is consistently over-, or underestimating the solubility of these compounds.

Diagram

Description automatically generated

Figure X: Chemical structures of the compounds found in the largest cluster of outliers, each containing at least one aromatic group.

References

(1) Spriewald, G.; Caswara, C.; Rodríguez-Guerra, J. *T005 · Compound clustering*. https://projects.volkamerlab.org/teachopencadd/talktorials/T005\_compound\_clustering.html.

(2) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics* **2015**, *7* (1). https://doi.org/10.1186/s13321-015-0069-3.

(3) Butina, D. Unsupervised Data Base Clustering Based on Daylight’s Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences* **1999**, *39* (4), 747–750. https://doi.org/10.1021/ci9803381.