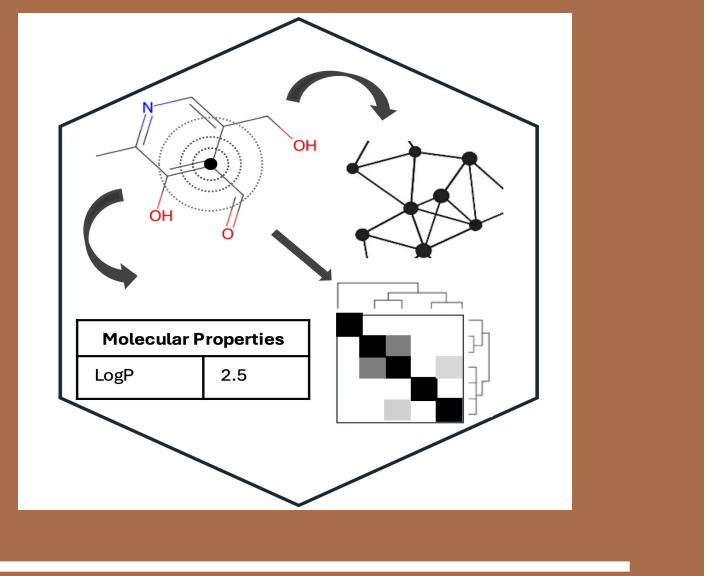
- **Informatics: Metabolomics, Lipidomics, and Glycomics**

CHEMPASS: A Tool for Rapid Chemical Structure Similarity Scoring and Clustering from Common Identifiers

Kruttika Dabke[1], Shiva Patre[1], Darren Kessner[1], Jonathan E. Katz[1,2]

[1] Ellison Medical Institute, Los Angeles, CA
[2] Department of Medicine, Keck School of Medicine,
University of Southern California, Los Angeles, CA
Correspondence: jkatz@emila.org
Reprints: https://reprints.katzlab.org/



Introduction

This project emerged from the need to build a set of internal standards for exposure studies. We wanted "class representatives" for common compounds of exposure as well as common endogenous compounds. Given that these standards were meant to be included with every sample, minimizing their number was crucial for cost and to simplify handling. To achieve this, we sought to create a "class based" exposure standard set where representative compounds could serve as surrogates for chemically similar compounds (e.g., Valine as a "stand-in" for amino acids like Alanine, Leucine, or Isoleucine).

To produce our list we started with a compound superset, generated molecular fingerprints using Extended Connectivity Fingerprinting (ECFP) with a radius of 4 atoms and then clustered the compounds based on Tanimoto similarity.

While doing this analysis, we realized that many existing tools that generate molecular fingerprints and compute chemical similarity require expertise, limiting accessibility. So, we developed **CHEMPASS**, a user-friendly web application that accepts common chemical identifiers (or SMILES), generates molecular fingerprints, and organizes compounds into clusters based on similarity scores.

We hope users find this app as useful as we have – from designing internal standard panels through optimizing drug efficacy studies to managing compound libraries.

Methods

The application asks the user to provide compound identifiers in the form of CIDs/DTXSIDs/SMILES or a mixture of all three. If provided with CIDs/DTXSIDs, the app obtains SMILES for those identifiers from PubChem (via the PUG REST API). The SMILES are converted to fingerprints using the RDKit Python library. Molecular properties (described in the Lipinski's Rule of 5 among others) are pulled for each compound from PubChem. Pairwise Tanimoto similarity scores are calculated for all compounds. The distance matrix is calculated from the matrix of similarity scores and then used to cluster structurally similar compounds using Butina clustering. The distance matrix is also used for performing hierarchical clustering and non-metric multidimensional scaling (NMDS) to better visualize clustering of structurally similar compounds.

The web application was developed using the R Shiny framework with Reticulate package, enabling interoperability between R and Python. Deployment is automated via GitHub Actions, which builds a Docker container using a Dockerfile generated from a standardized Cookiecutter template. The application is hosted securely with an SSL certificate issued by Certbot.

Results

The web application engages in a 3 step process (**Figure 2**):

Step 1 - Upload file: The user can input a mixture of compound identifiers in the form of CIDs/DTXSIDs/SMILES which get processed into a table linking the input IDs to SMILES obtained from PubChem along with their molecular properties (**Table 1**).

Bonus result!! A comprehensive CSV file is generated, containing the input chemical IDs and key descriptors based on Lipinski's Rule of Five. These descriptors—such as logP values, topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and molecular weight—are essential indicators of a drug's pharmacokinetics in the human body.

Step 2 - Generate Fingerprints: Next, the user selects a fingerprint type - either ECFP4 or Functional Class Fingerprints (FCFP4) from the drop down menu. Distance matrix calculated from the Tanimoto similarity matrix is used to render a heatmap within the app in the 'Heatmap' tab (**Figure 3.D**). The heatmap is generated using the ComplexHeatmap() package, with clustering method for rows and columns set to ward.D2.

Step 3 - Generate Butina Clusters: Finally, the user selects a clustering cut off from the drop down menu, which generates clusters of structurally similar compounds using the Butina algorithm (**Figure 4.A-B**). The clustered compounds are rendered in a PDF format under the 'Butina Clusters' tab and an interactive NMDS plot is generated in the 'NMDS Plot' tab. For Butina clusters, each cluster designates a central molecule—the one most structurally similar to all others in that group.

Designed with a user-friendly graphical interface, the application requires no coding expertise, making it accessible to a broad range of researchers. The application is hosted as a web-based platform with backend support for Python and R libraries to ensure ease of use and broad accessibility.

Conflict of Interest

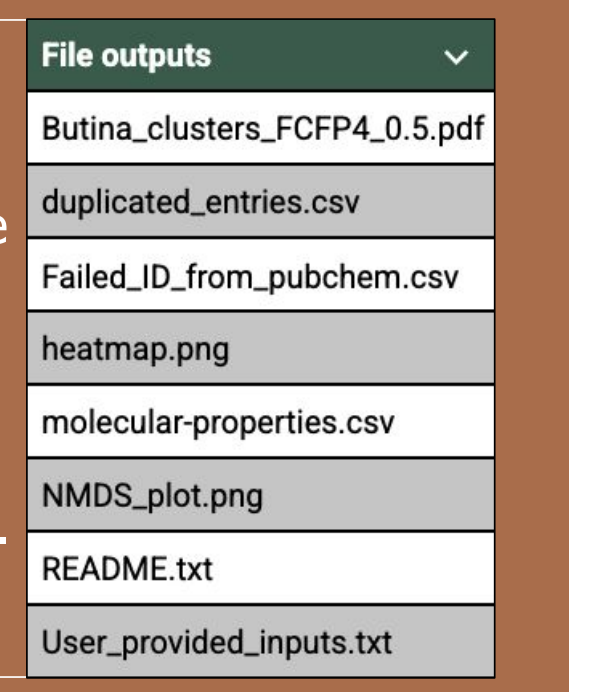
The authors declare no competing financial interests.

Acknowledgement

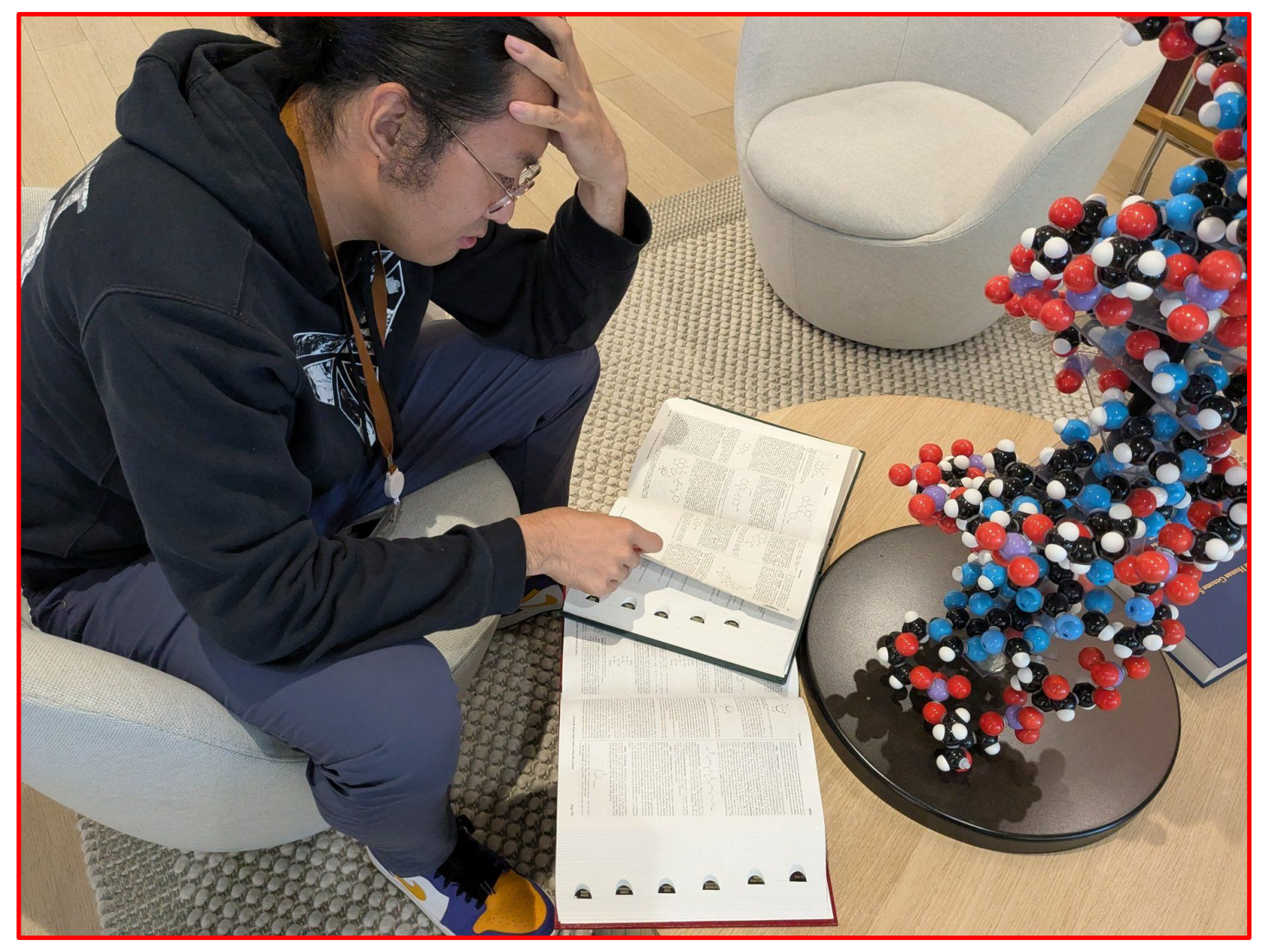
The authors would like to thank Warren Casey, Jennifer Kyle, John Wambaugh and Michelle Embry for their contributions in developing the exposure compound library and exposure reference standards, which led to the development of this web application.

Download Zip File

- The downloaded zip file contains a pdf of clusters as shown in **Figure 4.A-B**. The name of the pdf file captures the user defined fingerprint type and cluster cutoff.
- Duplicated entries and failed entries are stored in separate csv files and provided to user for further inspection.
- High resolution images (.png) for Heatmap (**Figure 3.D**) and NMDS plot (**Figure 4.C**) are provided for download.
- Molecular properties of compounds are provided in a separate csv file as shown in **Table 1**.
- README is provided with additional information on each of the downloaded files and some background information.
- User defined parameters are captured and stored in a text file and provided to users.



File outputs
Butina_clusters_FCFP4_0.5.pdf
duplicated_entries.csv
Failed_ID_from_pubchem.csv
heatmap.png
molecular-properties.csv
NMDS_plot.png
README.txt
User_provided_inputs.txt



Collating collections of compounds is a conundrum!

Getting physical properties of compounds is effort.
Grouping compounds by chemical similarity is effort.

Chempass solves both of these problems.

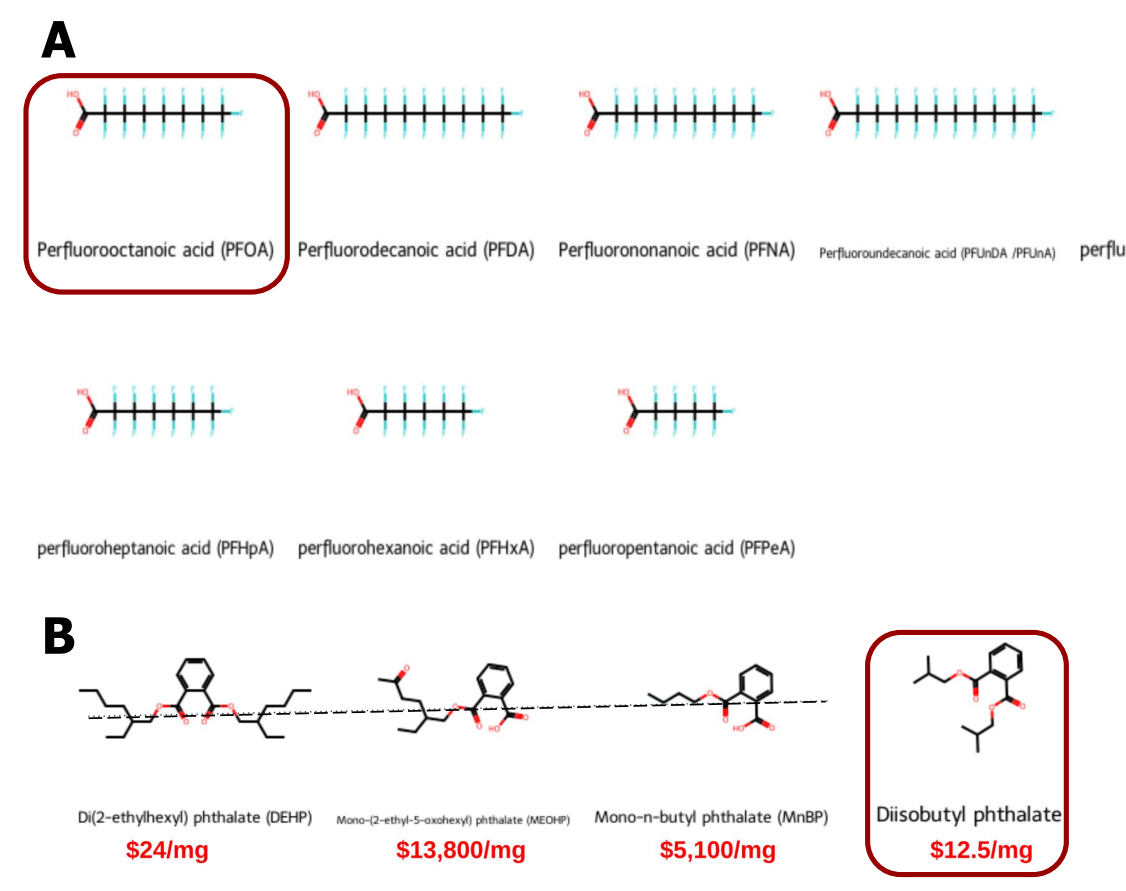

A
Perfluorooctanoic acid (PFOA)
Perfluorodecanoic acid (PFDA)
Perfluorononanoic acid (PFNA)
Perfluoroundecanoic acid (PFUnDA /PFUnA)
perfluorobutanoic acid (PFBA)
perfluoroheptanoic acid (PFHpA)
perfluorohexanoic acid (PFHxA)
perfluoropentanoic acid (PFPeA)
B
Di(2-ethylhexyl) phthalate (DEHP)
$24/mg
Mono-(2-ethyl-5-oxohexyl) phthalate (MEOHP)
$13,800/mg
Mono-n-butyl phthalate (MnBP)
$5,100/mg
Diisobutyl phthalate
$12.5/mg

Figure 1: Practical application:
We set out to create a set of internal standards for an exposure study with a list of ~300 compounds. Molecular fingerprinting and Butina clustering (distance cutoff = 0.5) helped reduce that list and either select "iconic" compounds from a cluster (**A**) or helped remove compounds belonging to the same cluster based on price or availability (**B**).

An interactive web app to:
Cluster Compounds by Structure Similarity,
and get their **Molecular Properties –**
no coding needed.

https://chempass.emilabs.org/

For code availability see:
https://reprints.katzlab.org/

Website Interface

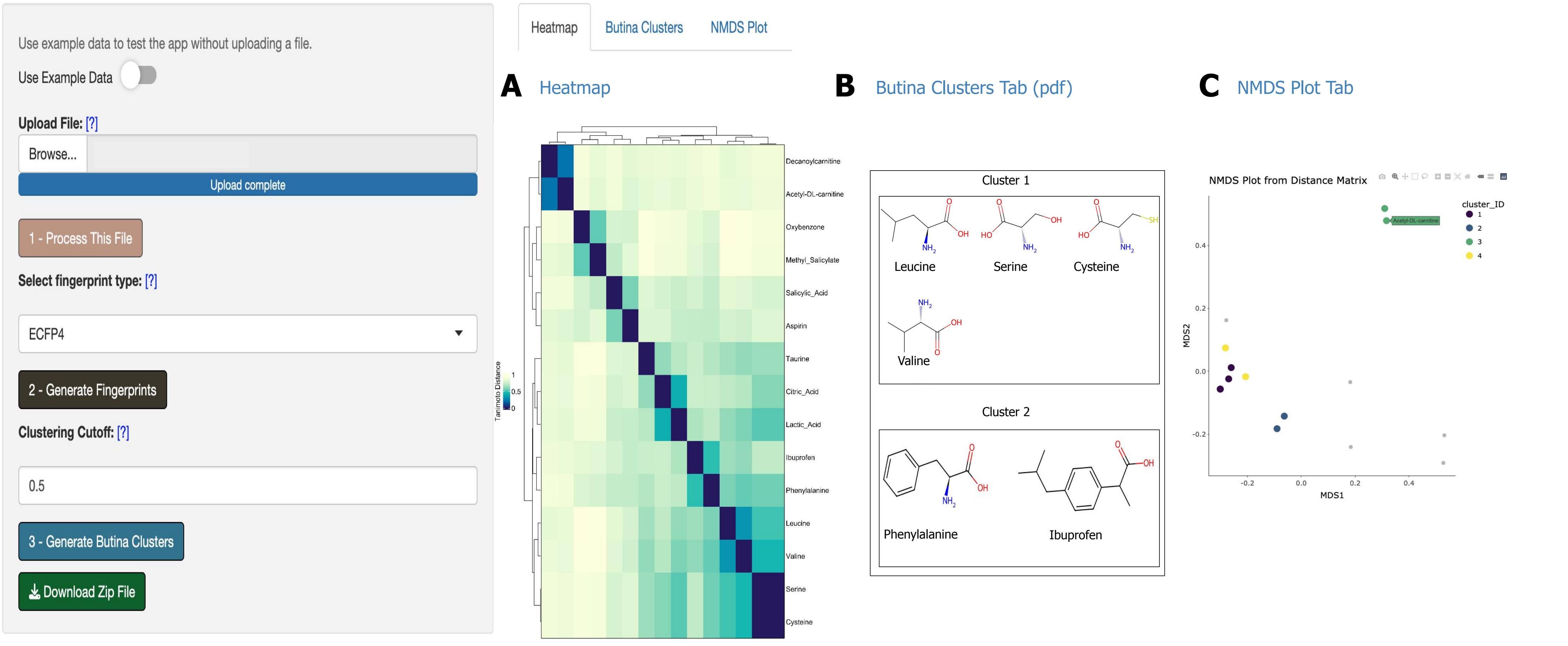
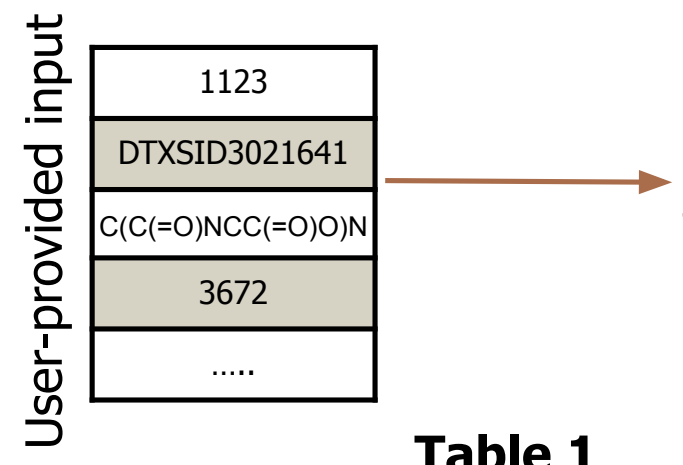


Figure 2: Website interface and outputs: Hierarchical clustering of a user-provided compound list (**A**) based on Tanimoto distance matrix, as displayed in the Heatmap tab, rendered upon clicking the "Generate Fingerprints" action button. Representative images of Butina clustering results (**B**) and Non-metric Multidimensional Scaling (**C**) based on Tanimoto distance matrix and user defined clustering cutoff, as displayed in the Butina Clusters tab, rendered upon clicking the "Generate Butina Clusters" action button. NMDS plot displays clustered compounds and is fully interactive within the web app.

Website Action Buttons:

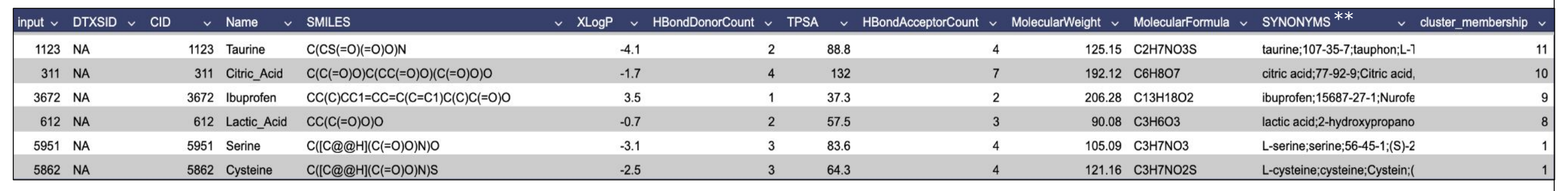
1- Process This File

Molecular properties are pulled from PubChem using PUG-REST API*; a web interface for accessing PubChem data and services.



```
requests.get(url)
url =
f'https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/{CIDs}/property/Title,SMILES,XLogP,HBondDonorCount,TPSA,HBondAcceptorCount,MolecularWeight,MolecularFormula/csv'
```

Table 1



input	DTXSID	CID	Name	SMILES	XLogP	HBondDonorCount	TPSA	HBondAcceptorCount	MolecularWeight	MolecularFormula	SYNONYMS **	cluster_membership
1123	NA	1123	Taurine	C(CS(=O)(=O)O)N	-4.1	2	88.8	4	125.15	C2H7NO3S	taurine;107-35-7;tauphon;L-1	11
311	NA	311	Citric_Acid	C(C(=O)O)C(CC(=O)O)(C(=O)O)O	-1.7	4	132	7	192.12	C6H8O7	citric acid;77-92-9;Citric acid,	10
3672	NA	3672	Ibuprofen	CC(C)CC1=CC=C(C=C1)C(C)C(=O)O	3.5	1	37.3	2	206.28	C13H18O2	ibuprofen;15687-27-1;Nurofe	9
612	NA	612	Lactic_Acid	CC(C(=O)O)O	-0.7	2	57.5	3	90.08	C3H6O3	lactic acid;2-hydroxypropano	8
5951	NA	5951	Serine	C([C@@H](C(=O)O)N)O	-3.1	3	83.6	4	105.09	C3H7NO3	L-serine;serine;56-45-1;(S)-2	1
5862	NA	5862	Cysteine	C([C@@H](C(=O)O)N)S	-2.5	3	64.3	4	121.16	C3H7NO2S	L-cysteine;cysteine;Cystein;(	1

User-provided input—whether as CID, DTXSID, or SMILES—is processed via PubChem to retrieve molecular properties, which are compiled in **Table 1**. This table is included in the downloadable zip file as a .csv. It allows users to easily access PubChem-derived molecular properties for evaluating the oral bioavailability of their compounds.

**SYNONYMS column contains up to 10 synonyms for a molecule listed in PubChem separated by ";".

*Kim, S, Thiessen, PA, Cheng, T, Yu, B & Bolton, E. E. Nucleic Acids Res. 46, W563–W570 (2018)

2- Generate Fingerprints

(+ generate Tanimoto Similarity scores and distance heatmaps)

Morgan fingerprints are computed from the SMILES (representations of user-submitted compounds).+

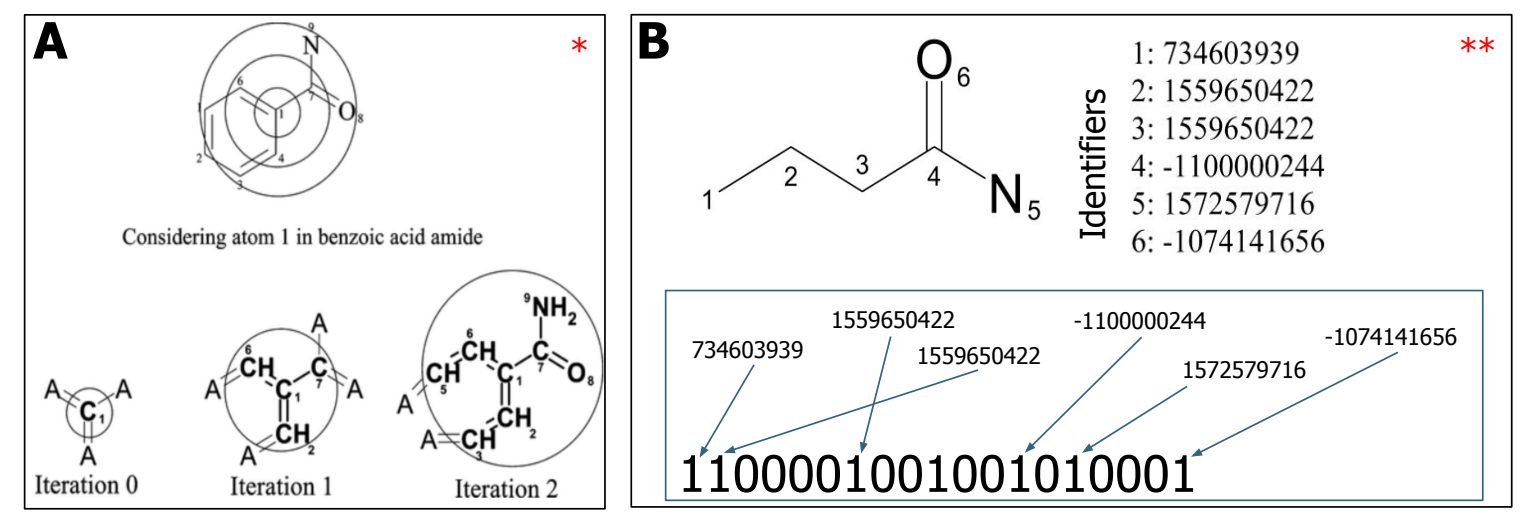


Figure 3: Extended Connectivity Fingerprinting (ECFP). Each atom in a compound was iteratively accessed at a diameter of 4 neighbors (**A**) to produce a unique binary representation representing that structure (**B**). The totality of representations were collapsed into a 2048 bit space creating the final hash for that molecule (or"fingerprint").

C

Tanimoto coefficient:

$$S(A,B) = \frac{c}{a + b + c}$$

- a is the number of 'on' bits in molecule A.
- b is number of 'on' bits in molecule B.
- c is the number of bits that are 'on' in both molecules.

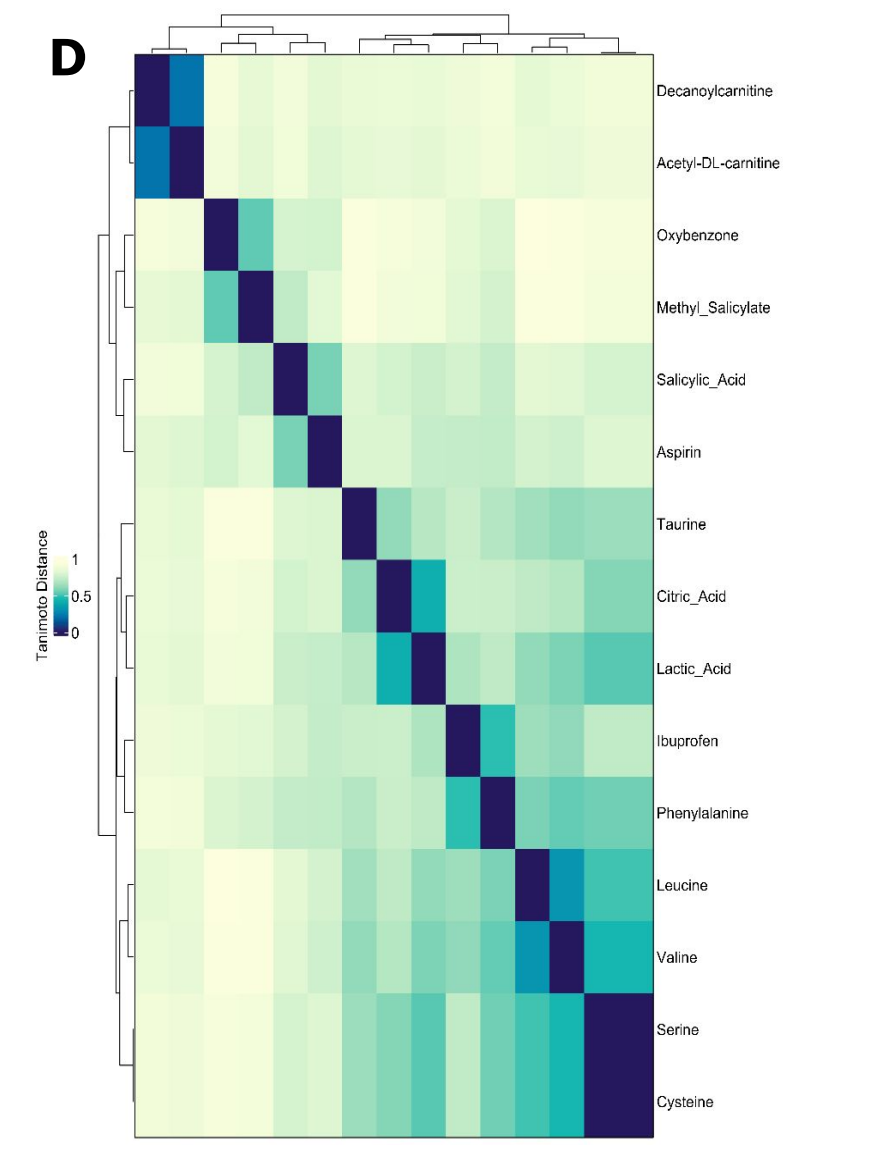


Figure 3: Visualization of Tanimoto Distance Matrix:
- Tanimoto similarity scores were computed for all pairs of compound fingerprints (**C**).
- Scores range from 0 (no similarity) to 1 (identical structures), with higher values indicating greater structural similarity.
- The dissimilarity matrix or distance matrix is calculated from the similarity matrix and used for clustering compounds with Hierarchical clustering (**D**), Butina clustering and NMDS plots.

*J. Chem. Inf. Model. 2010, 50, 5, 742–754
**https://towardsdatascience.com/a-practical-introduction-to-the-use-of-molecular-fingerprints-in-drug-discovery-7f15021be2b1/
+RDKit: Open-source cheminformatics. https://www.rdkit.org

3- Generate Butina Clusters

Butina clustering is performed using the Tanimoto distance matrix generated in step 2 and user defined clustering cutoff.

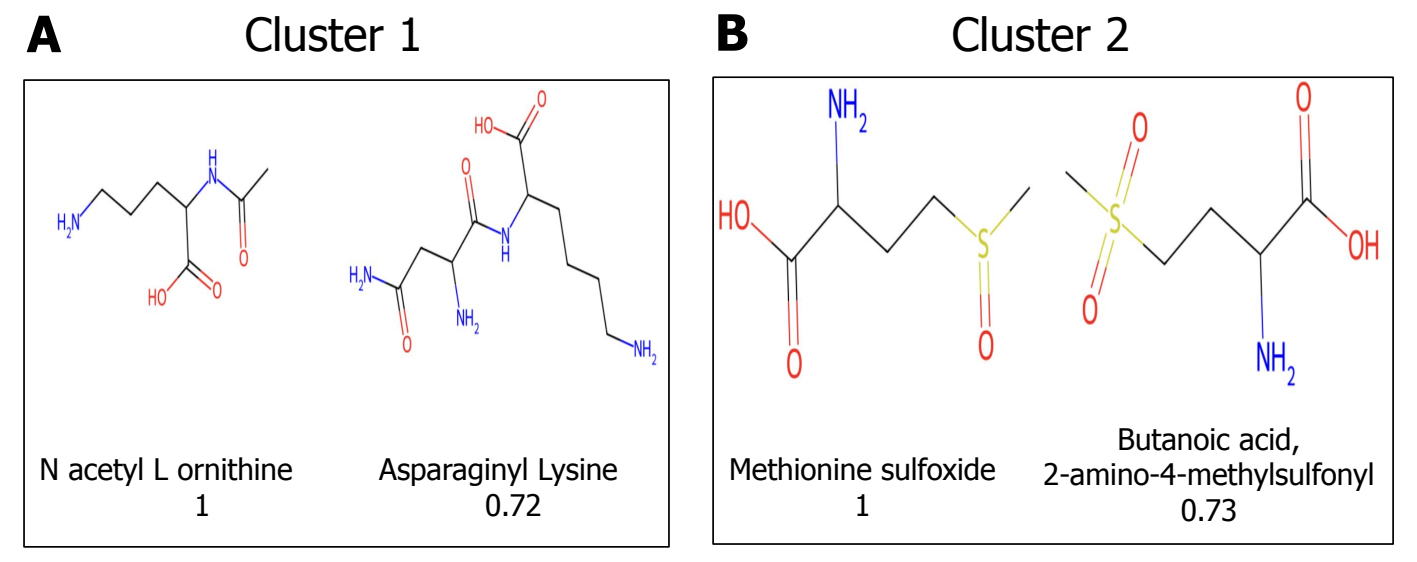


Figure 4: Butina Clustering: Compounds are grouped based on Tanimoto distance using a user-defined clustering cutoff. Each cluster is built around a *centroid*—the molecule with the most neighbors. Assigned compounds are excluded from forming or joining other clusters. The process repeats until all compounds are clustered or labeled as *singletons.**

Cluster Report Format: Each cluster is shown on a separate PDF page (**A-B,** cluster cutoff = 0.3). The cluster center appears first, followed by members and their similarity scores to the center (typically ≥ 1 – cutoff). Singletons are displayed one per page.

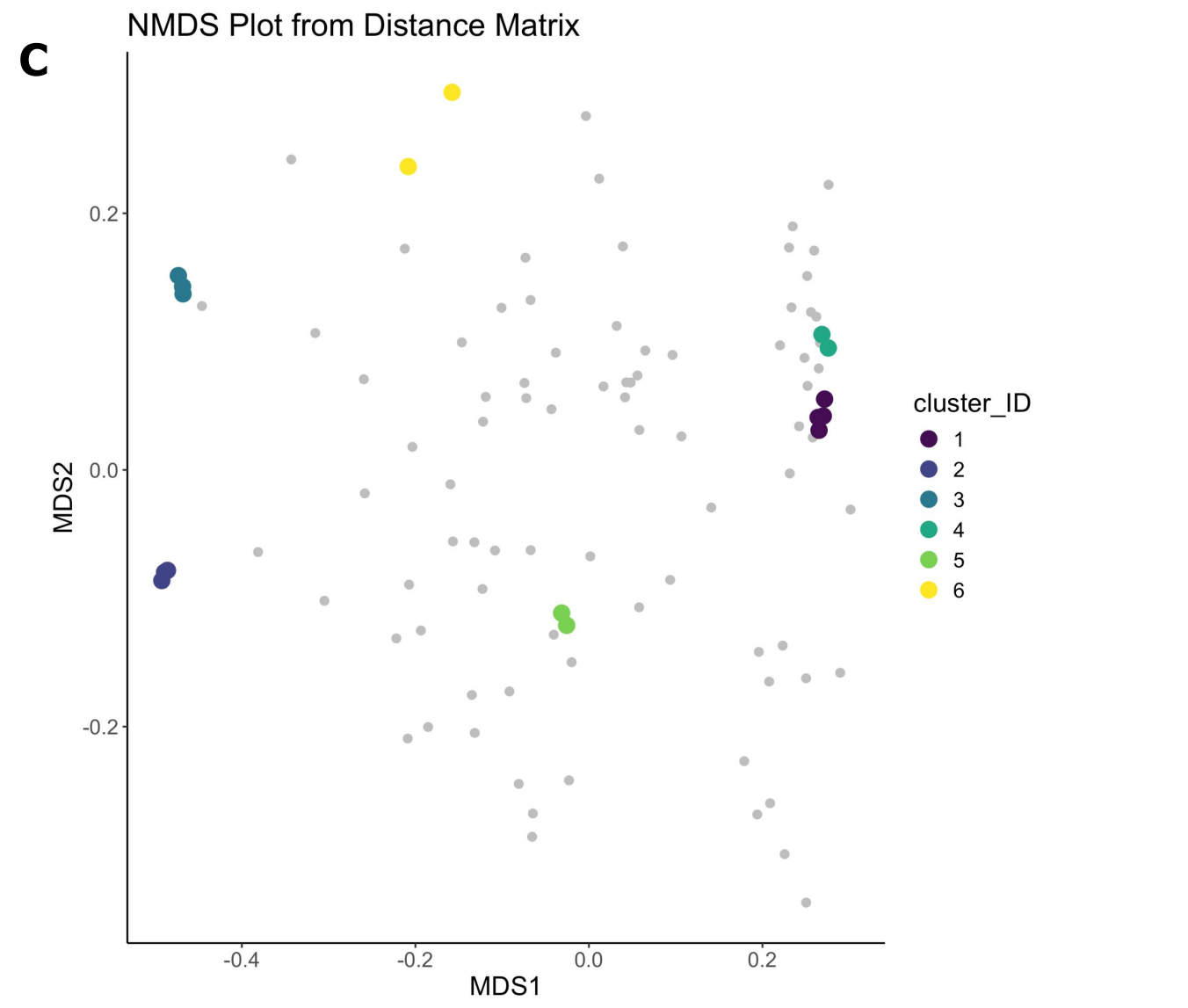


Figure 4: Non-metric Multidimensional Scaling (NMDS):
- NMDS is a flexible, distance-based visualization method that preserves the *rank order* of pairwise distances (**C**), not the actual distances themselves.**
- While clustering groups compounds based on structural similarity using discrete thresholds, NMDS provides a continuous, visual overview of how compounds relate to one another in low-dimensional space.

*https://projects.volkamerlab.org/teachopencadd/talktorials/T005_compound_clustering.html
**https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/nmds/

Ellison
Medical
Institute