Optimization of Clustering Algorithms for Grouping of Complex Chemical Substances Based on Chemical and Biological Characteristics

Melis Onel[1], Burcu Beykal[1], Fabian A. Grimm[2], Lan Zhou[3], Fred A. Wright[4], Ivan Rusyn[2], Efstratios N. Pistikopoulos[1]

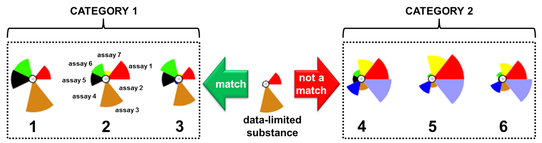
[1]Artie McFerrin Department of Chemical Engineering and Texas A&M Energy Institute, Texas A&M University, College Station, TX

[2]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX; [3]Department of Statistics, Texas A&M University, College Station, TX

[4]Bioinformatics Research Center, North Carolina State University, Raleigh, NC

Rationale

- **Texas A&M Superfund program goal:** To develop comprehensive tools and models for addressing exposure to chemical mixtures during environmental emergency-related contamination events.
- **Aim:** To design a **framework for optimal grouping of chemical mixtures based on their chemical characteristics and bioactivity properties**, and facilitate comparative assessment of their human health impacts through read-across.
- **Approach:** In order to explore the most optimal clustering algorithms that may be used to establish the chemical and biological similarity between complex substances or mixtures, we used several recent examples of chemical substances of Unknown or Variable composition Complex reaction products, and Biological materials (UVCB substances). UVCBs present a major challenge for registrations under the REACH and US High Production Volume regulatory programs. In addition to frequent variations in their chemical composition, **many gaps in available toxicity data preclude confident categorization of these substances for read across applications.**

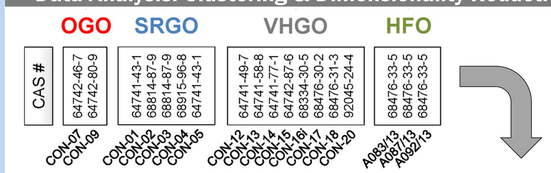


- Here, we present a comprehensive computational approach using different clustering algorithms to categorize UVCBs according to global similarities in a) their chemical composition using Gas Chromatography-Gas Chromatography Flame Ionization Detector (GCXGC-FID) and Ion Mobility Mass Spectrometry (IM-MS) [1], b) their bioactivities from in vitro screening in human cells [2].

Motivation: Categorization of high-dimensional analytical data by exploring different clustering algorithms and dimensionality reduction techniques.

Data Integration & Processing

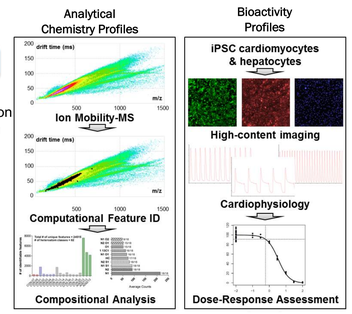
Data Analysis: Clustering & Dimensionality Reduction



Hierarchical Clustering

- Kendall Correlation
- Spearman Correlation
- Pearson Correlation

- Ward's Method
- Average Linkage
- Complete Linkage

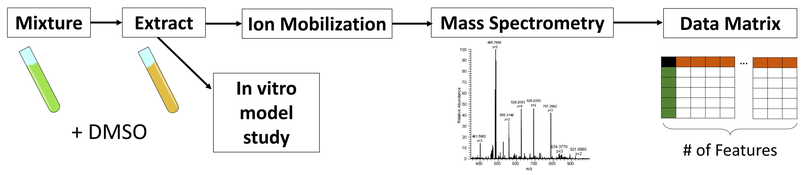


- Hierarchical Clustering • 9 Clustering Methods Explored
- Principle Component Analysis • Dimensionality Reduction
- Fowlkes–Mallows Index • Comparative Analysis

➢ The clustering algorithms are evaluated with the **Fowlkes–Mallows index** in order to provide a comparative assessment for the clustering quality.

Analytical Chemistry Profiles

Data Acquisition: Targeted & Untargeted Analysis

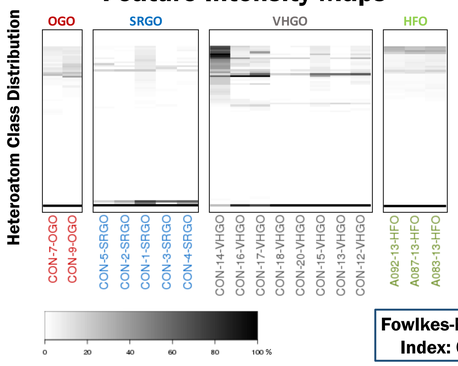


Mass Spectrometry and Ion Mobilization Devices at Texas A&M

Chemical Fingerprinting & Cluster Analysis of UVCBs

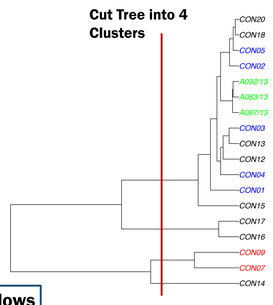
IM-MS- 82 Features — **Average Linkage – Pearson Correlation**

Feature Intensity Maps

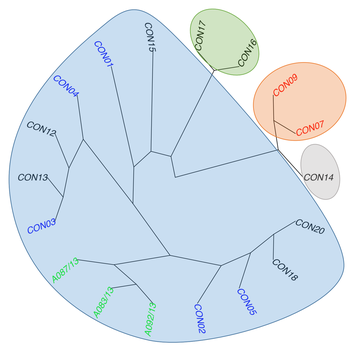


Fowlkes-Mallows Index: 0.431

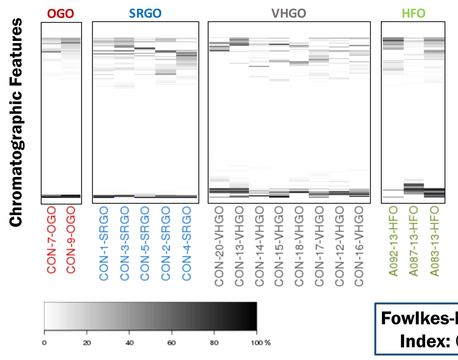
Dendrograms



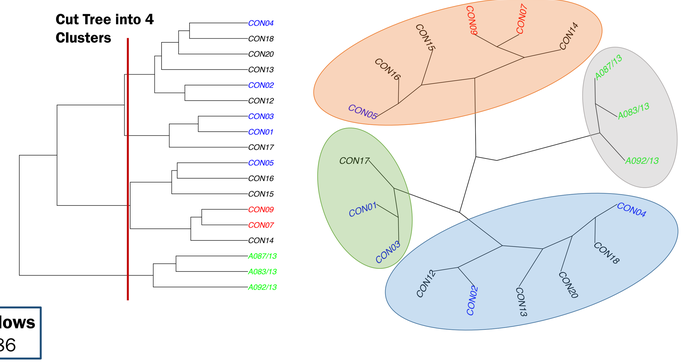
Unrooted Trees



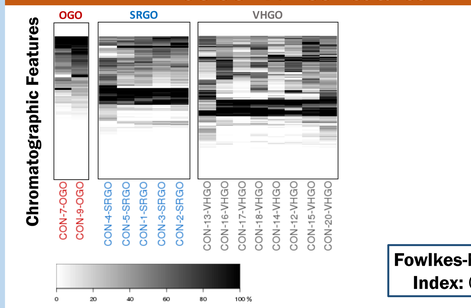
IM-MS- 445 Features — **Ward's Method – Spearman Correlation**



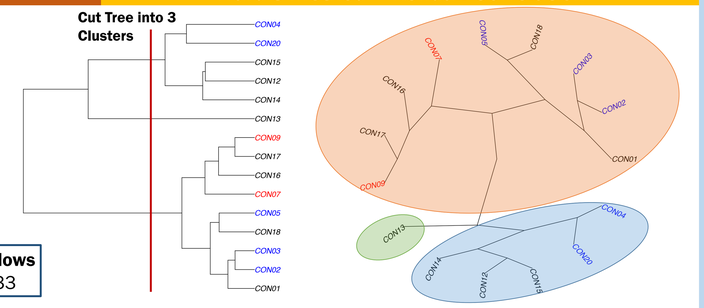
Fowlkes-Mallows Index: 0.386



GC-GC-FID 260 Features — **Ward's Method – Kendall Correlation**

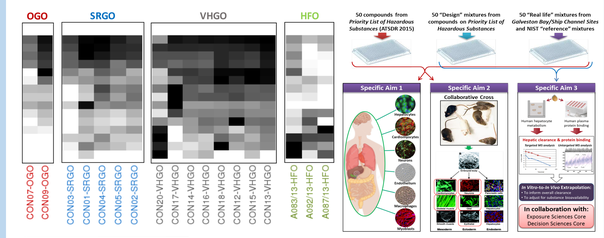


Fowlkes-Mallows Index: 0.333

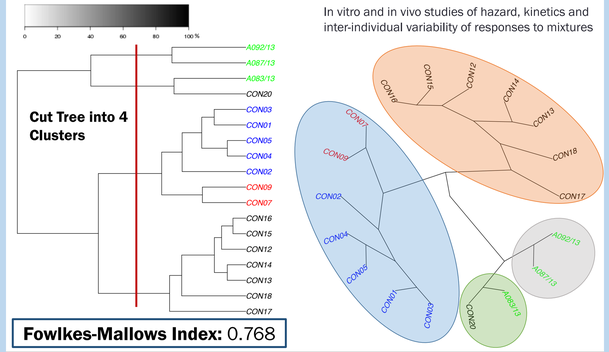


Bioactivity Profiles

Data Acquisition & Analysis

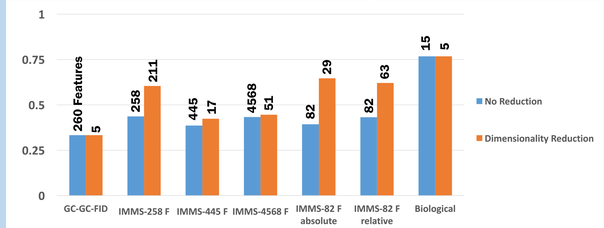


In vitro and in vivo studies of hazard, kinetics and inter-individual variability of responses to mixtures



Fowlkes-Mallows Index: 0.768

Effect of Dimensionality Reduction



Conclusions

- Prototypical high-production volume UVCBs, can be **categorized using global similarities** in their physico-chemical descriptors, **global compositional analysis** using Ion Mobility-Mass Spectrometry, and **bioactivity profiles** using multi-parametric HCS of iPSC-derived cell types.
- Fowlkes-Mallows index has been implemented to **optimize clustering algorithms**.

Future Work

- A combinatorial approach using quantitative chemical analysis, high-content *in vitro* screenings, and **subsequent computational data integration and visualization** will be performed which is **anticipated to improve chemical-biological read-across applications**.
- Other dimensionality reduction techniques will be tested to **further improve the framework**.

Key References

- Grimm FA et. al. "A chemical–biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives". Green Chemistry, 2016, 18, 4407.
- Grimm FA et. al. "Grouping of Petroleum Substances as Example UVCBs by Ion Mobility-Mass Spectrometry to Enable Chemical Composition-Based Read Across". Environmental Science & Technology. 2017, 51 (12), pp 7197–7207
- Onel, M.; Kieslich, C. A.; Guzman, Y. A.; Floudas, C. A.; Pistikopoulos, E. N. "Big Data Approach to Batch Process Monitoring: Simultaneous Fault Detection and Identification Using Nonlinear Support Vector Machine-based Feature Selection". Computers & Chemical Engineering, 2017. (Under Review).
- Fowlkes, E. B.; Mallows, C. L. "A Method for Comparing Two Hierarchical Clusterings". *Journal of the American Statistical Association*. 1983, **78** (383): 553.

Funding is acknowledged from the NIEHS Superfund Research Program (P42-ES027704). The portions of the work were performed at the Texas A&M High-Performance Supercomputing Facility.