Towards characterizing the galaxies of biosolids chemical classes across the chemical universe

Paul Kruse[1,2], Caroline Ring[1]

1. Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA 2. Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, TN, USA

Abstract number: 4644

Paul Kruse | kruse.paul@epa.gov | ORCID: 0000-0001-5516-9717

Introduction

- Biosolids (treated sewage sludge) are applied to land
- Need for risk-based screening & prioritization of biosolids chemical contaminants — but data gaps make it difficult
- **Machine learning model to predict chemical concentrations in biosolids**
  - **Training data**: National Sewage Sludge Survey (NSSS) monitoring data (744 chemicals)
  - **Prediction set** (for example): TSCA Inventory (68k chemicals) [1]
- **Domain of applicability: How well does NSSS chemical space represent TSCA chemical space?**

Methods

- Chemical space characterized using **ClassyFire & ChemOnt** [2]
  - Structure-based classification
  - "Tree of life" hierarchical ontology
- Visualize chemical space using **tree-based visualizations** [3-10]
- **Quantify & visualize similarity** of training and prediction sets
  - Calculate established **similarity measures** for tree-based ontologies [11-15]
    - Use ClassyFire classification as a "fingerprint" instead of structural fingerprint (Figure 2)
  - **Heatmap visualization** of similarity [16]

Discussion

- **NSSS is a fairly-representative subset of TSCA** (Figure 1)
- **NSSS is as similar to TSCA as TSCA is to itself**, and more similar than occurs by chance (Table 1)
- Heatmap: Identify *which* classes are better/worse represented (Figure 3)

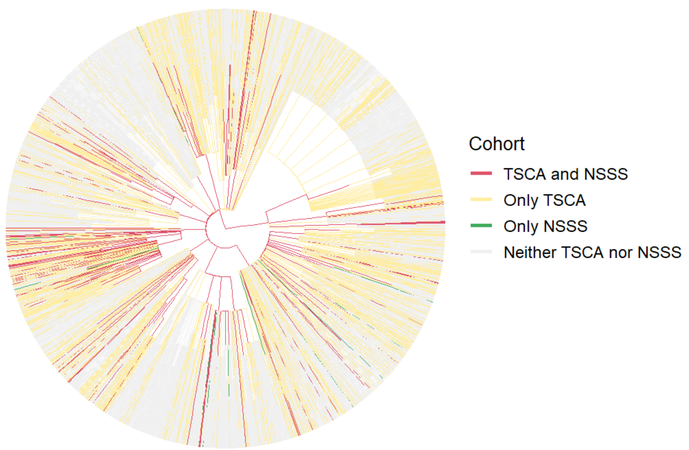

Figure 1. Full ChemOnt tree. Branches color-coded by representation in TSCA, NSSS, both, or neither

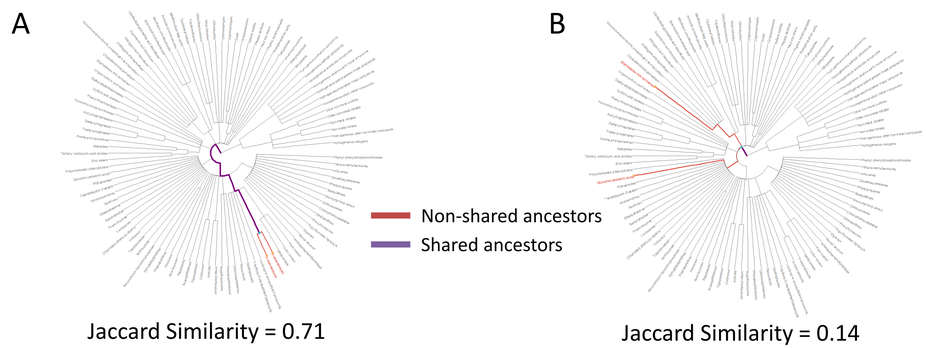

Figure 2. Jaccard similarity of ClassyFire classifications

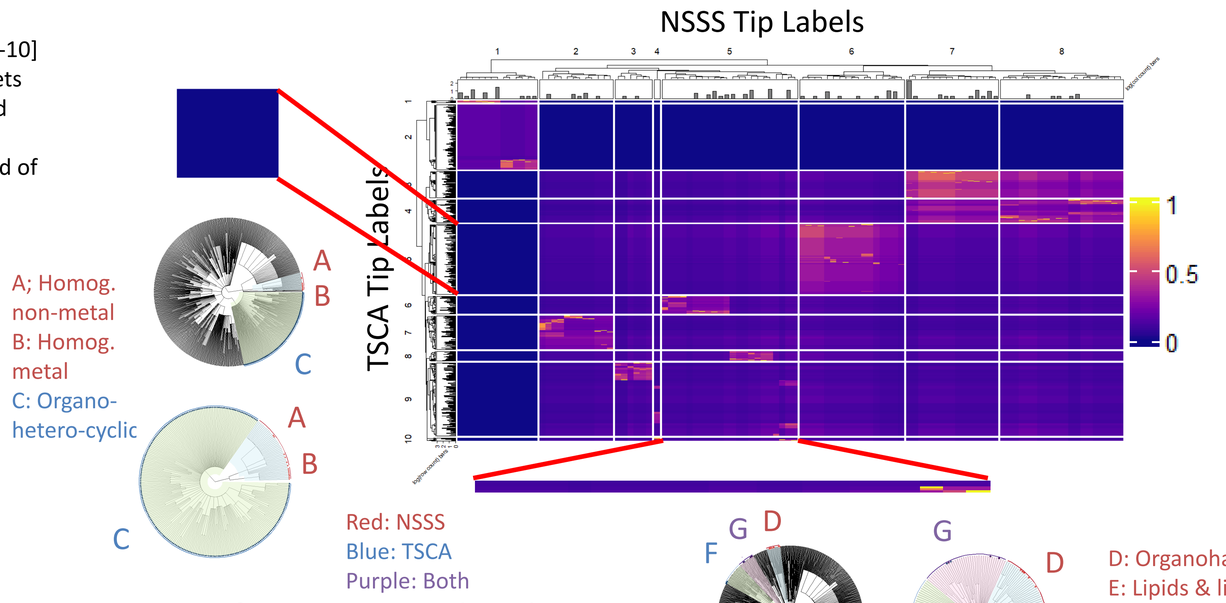

Figure 3. Heatmap of pairwise Jaccard similarities for NSSS vs. TSCA tip labels. Details: Chemical classes represented by two heatmap blocks.

Table 1. Average pairwise Jaccard similarity of labels in Tree1 vs. Tree2. Random trees: average over n = 100 random trees.

Tree1	Tree2	Jaccard
TSCA	TSCA	0.11
NSSS	NSSS	0.16
TSCA	NSSS	0.12
Random "TSCA"	Random "TSCA"	0.12
Random "NSSS"	Random "NSSS"	0.14
TSCA	Random "NSSS"	0.11
NSSS	Random "TSCA"	0.11

References