United States Environmental Protection Agency

Revisiting and updating chemical groupings with new approach methodologies

US EPA in collaboration with Health Canada, Environment Climate Change Canada

ACS Fall 2020 Virtual Meeting & Expo
Computational Strategies in Modern Agrochemical Discovery and De-risking Symposium
August 17 – 20, 2020

Office of Research and Development
Center for Computational Toxicology and Exposure

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. This presentation has not been reviewed for policy and is not for distribution.

Team members

Accelerating the Pace of Chemical Risk Assessment (APCRA)

- US EPA
  - Dan Chang
  - Kellie Fay
  - Kristan Markey
  - Martin Phillips
  - Grace Patlewicz
  - Ann Richard
  - Gino Scarano
  - Mahmoud Shobair
  - Ryan Lougee
  - Ellery Saluck (summer intern)
- Environment & Climate Change Canada (ECCC)
  - John Prindiville
  - Cristina Inglis
- Health Canada
  - Mark Lewis
- ILS
  - Kamel Mansouri

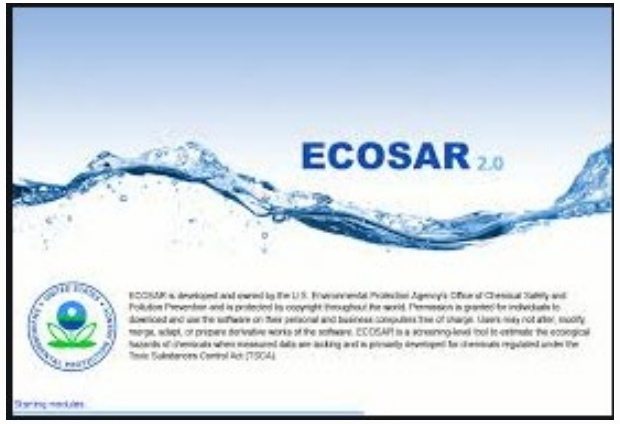
Overview

A chemical category is a group of chemicals whose physicochemical and human health and/or ecotoxicological properties and/or environmental fate properties are likely to be similar or follow a regular pattern, usually as a result of structural similarity. - OECD

Applications of chemical categorization include first tier assessment efforts and read across from structurally similar analogs:

- Toxic Substances Control Act (TSCA) New Chemical Program Chemical Categories (NCC; US EPA)
- ECOSAR (focus of presented work)

US EPA ECOSAR chemical classifications



- Class-based SAR to predict aquatic toxicity
- Classification scheme identifies excess toxicity
- Estimates **acute** and **chronic toxicity** based on accumulated data and past decisional precedents

Acute Effects:	Chronic Effects:
Fish 96-hr LC_{50}	Fish ChV
Daphnid 48-hr EC_{50}	Daphnid ChV
Algae 72/96-hr EC_{50}	Algae ChV

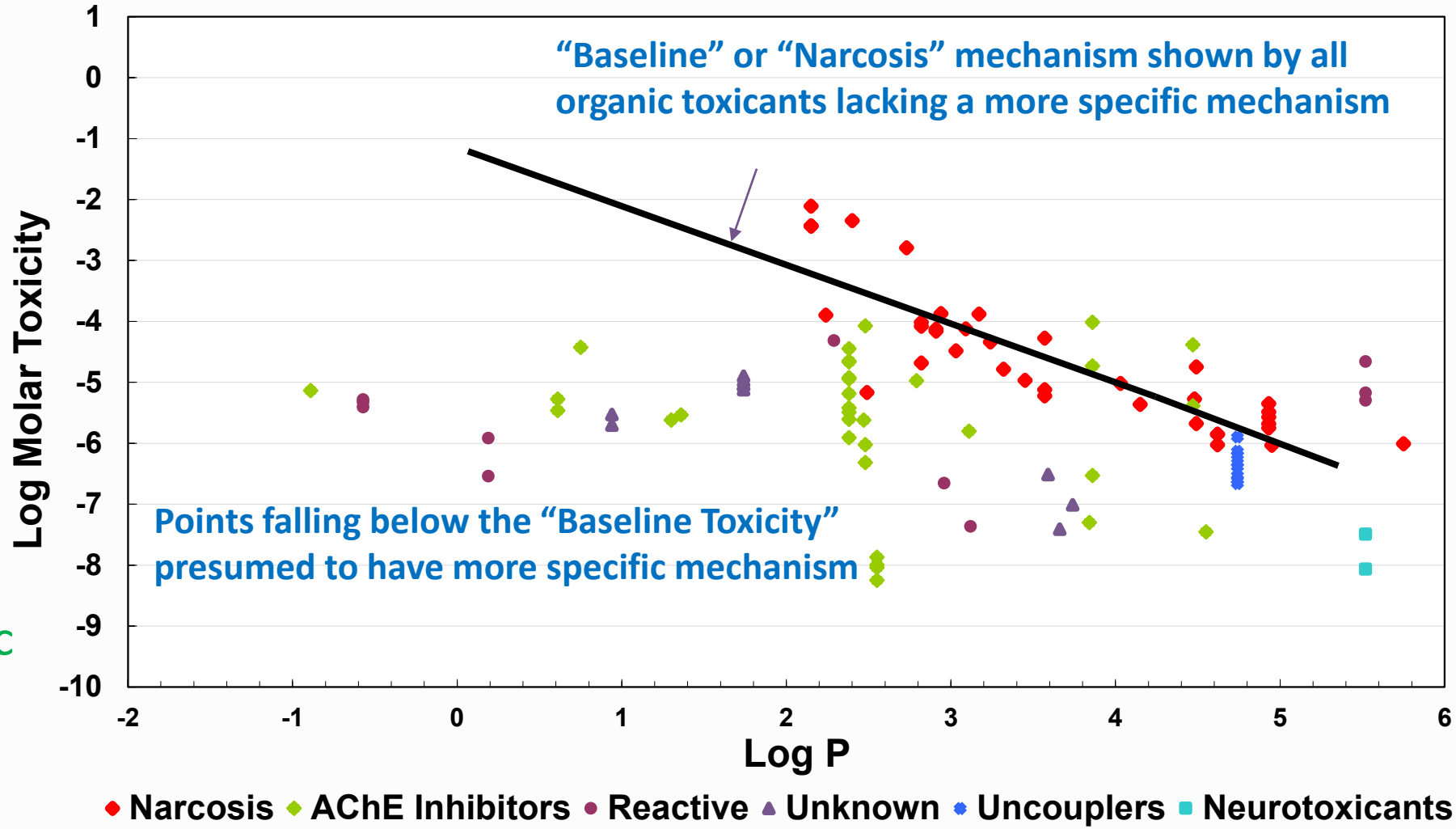
- Profiler in OECD QSAR Toolbox

Narcosis vs. specific-acting toxicity MOA

Less Toxic

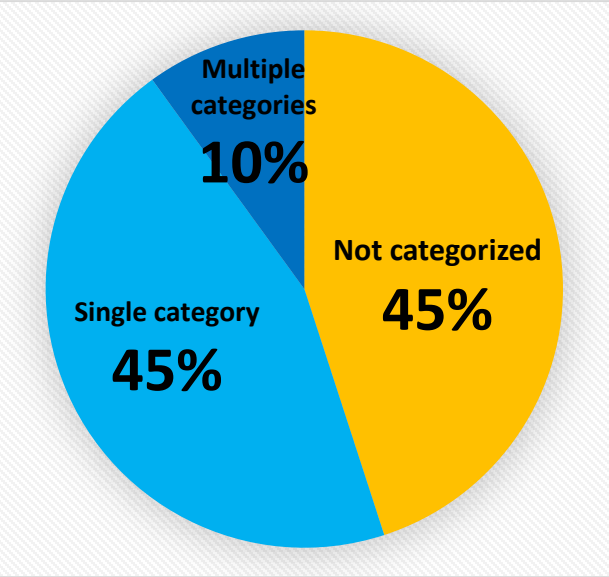
Regulators (ECCC) consider MOA information to determine the size of assessment factors

More Toxic



Office of Research and Development
Center for Computational Toxicology and Exposure

Potential approach for updating chemical categories

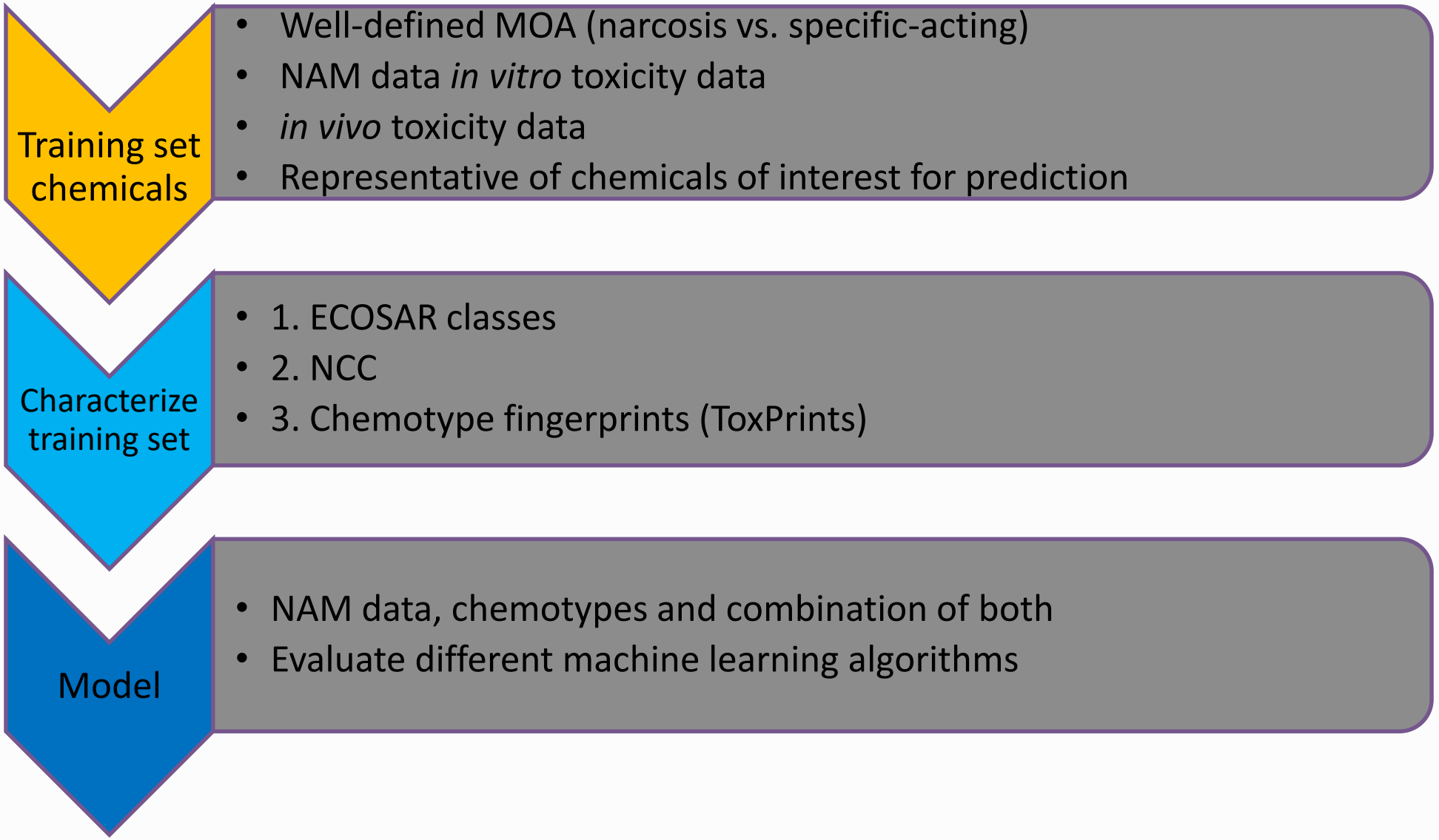


- Almost half of all New Chemical inventories across regulatory jurisdictions cannot be categorized using NCC or ECOSAR
- Some fall into multiple categories

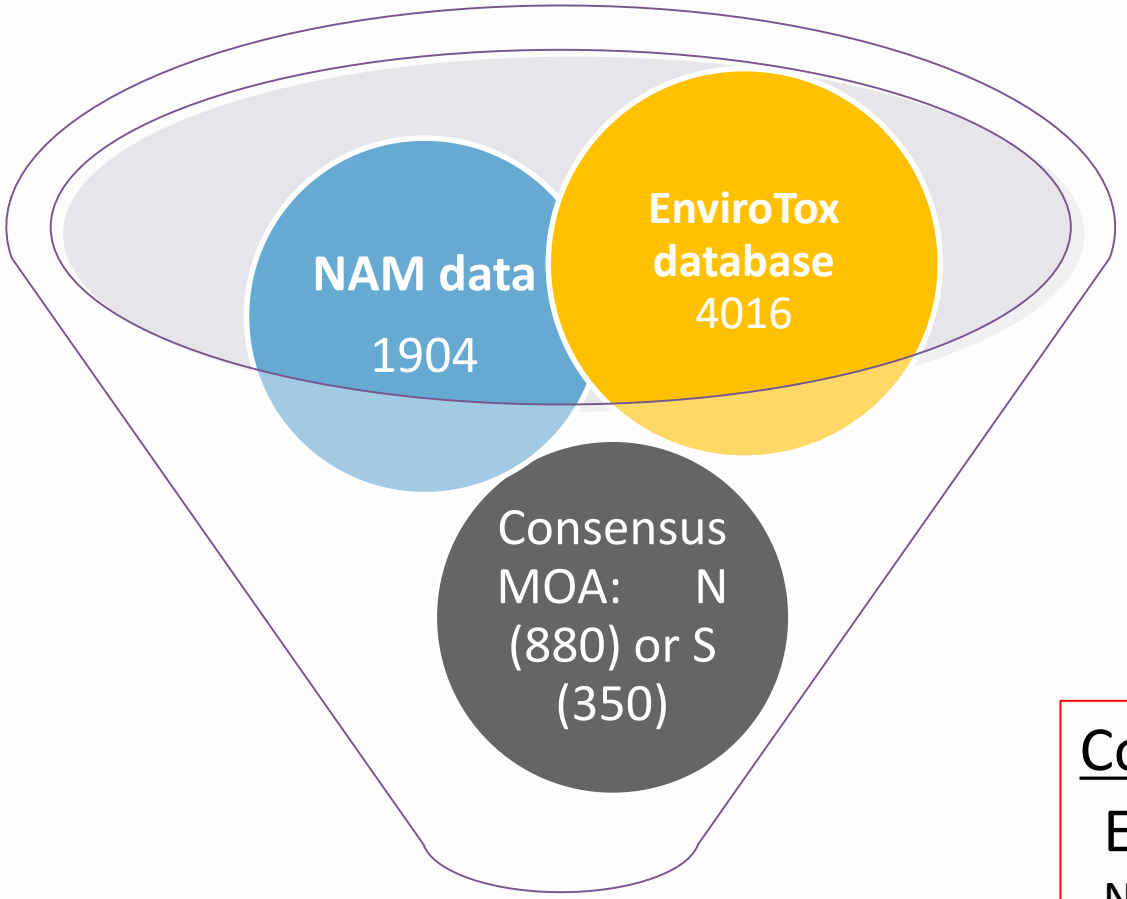
How do we update?

- Incorporate New Approach Methodologies (NAMs) – *i.e.*, ToxCast and Tox21 biological activity information
- Apply cheminformatic approaches

General approach



EnviroTox training set chemicals



Training set chemicals

1. Chemicals with *in vivo* eco-data – from the EnviroTox[1] database – 4016
2. Sub-selection for chemicals with NAM data (ToxCast and Tox21) - 1904
3. MOA predictions based on 4 publicly-available classification models
   - VERHAAR, ASTER, OASIS, TEST
   - Each predicts Narcotic, Specific-Acting or Unclassified

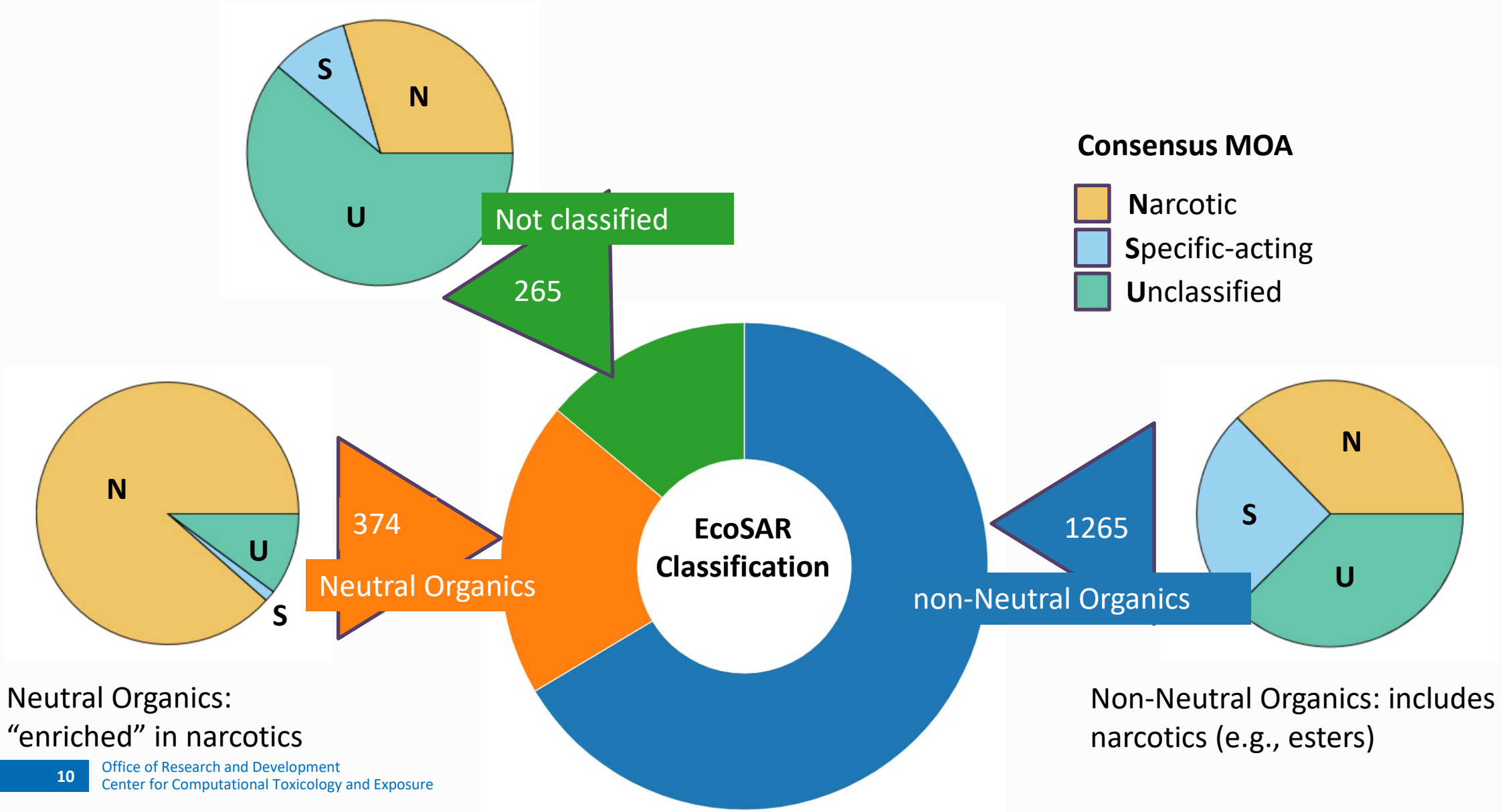
Consensus MOA with confidence scores[2]

Examples:	Results:
NNNN = N, score =3	880 Narcotic
NNSN = N, score= 2	350 Specific-acting
SUSS = S, score= 2	674 Unclassified
NUNS = U, score = 0	

[1]Health and Environmental Sciences Institute (HESI). 2019. EnviroTox Database & Tools. Version 1.1.0 Available: http://www.envirotoxdatabase.org/

[2] Kienzler et al.. Environ Toxicol and Chem. 2019, 38(10) 2294-2304

Characterize EnviroTox training set chemicals: ECOSAR classes

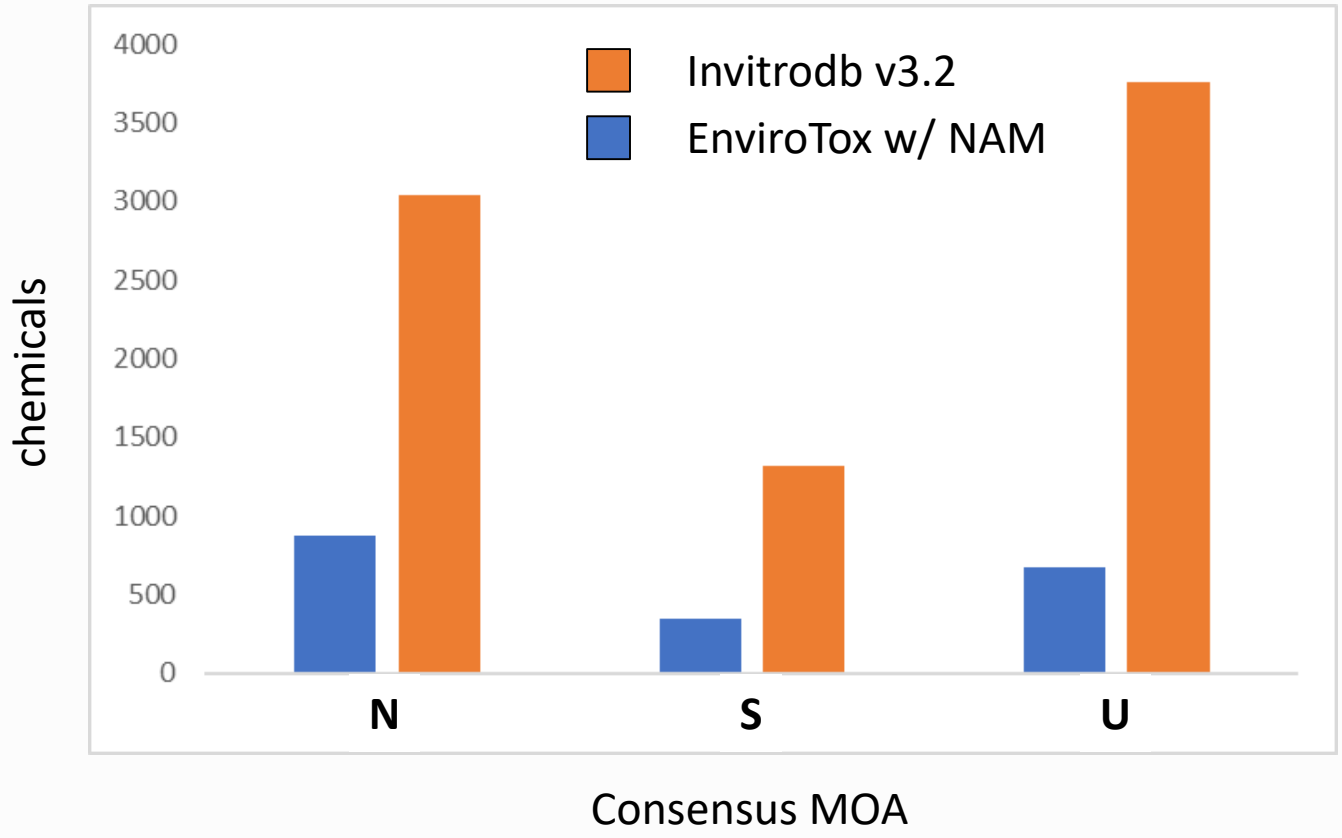


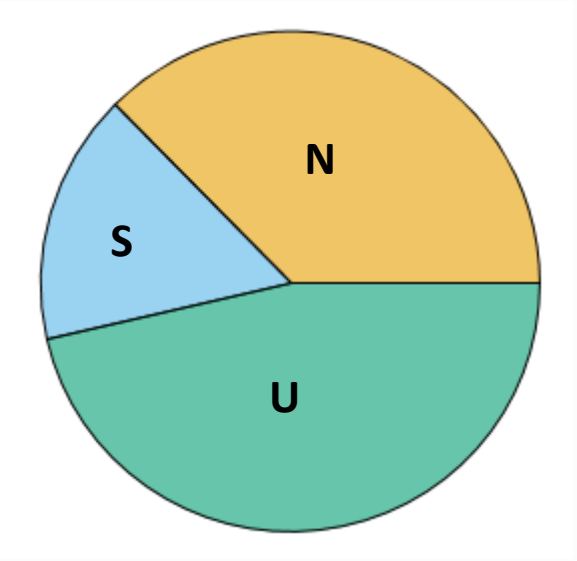
Neutral Organics:
"enriched" in narcotics

Non-Neutral Organics: includes narcotics (e.g., esters)

Expanding the chemical space of the EnviroTox dataset

- Added 6215 chemicals with NAM data (invitrodb v3.2)
- Applied the same consensus MOA methodology





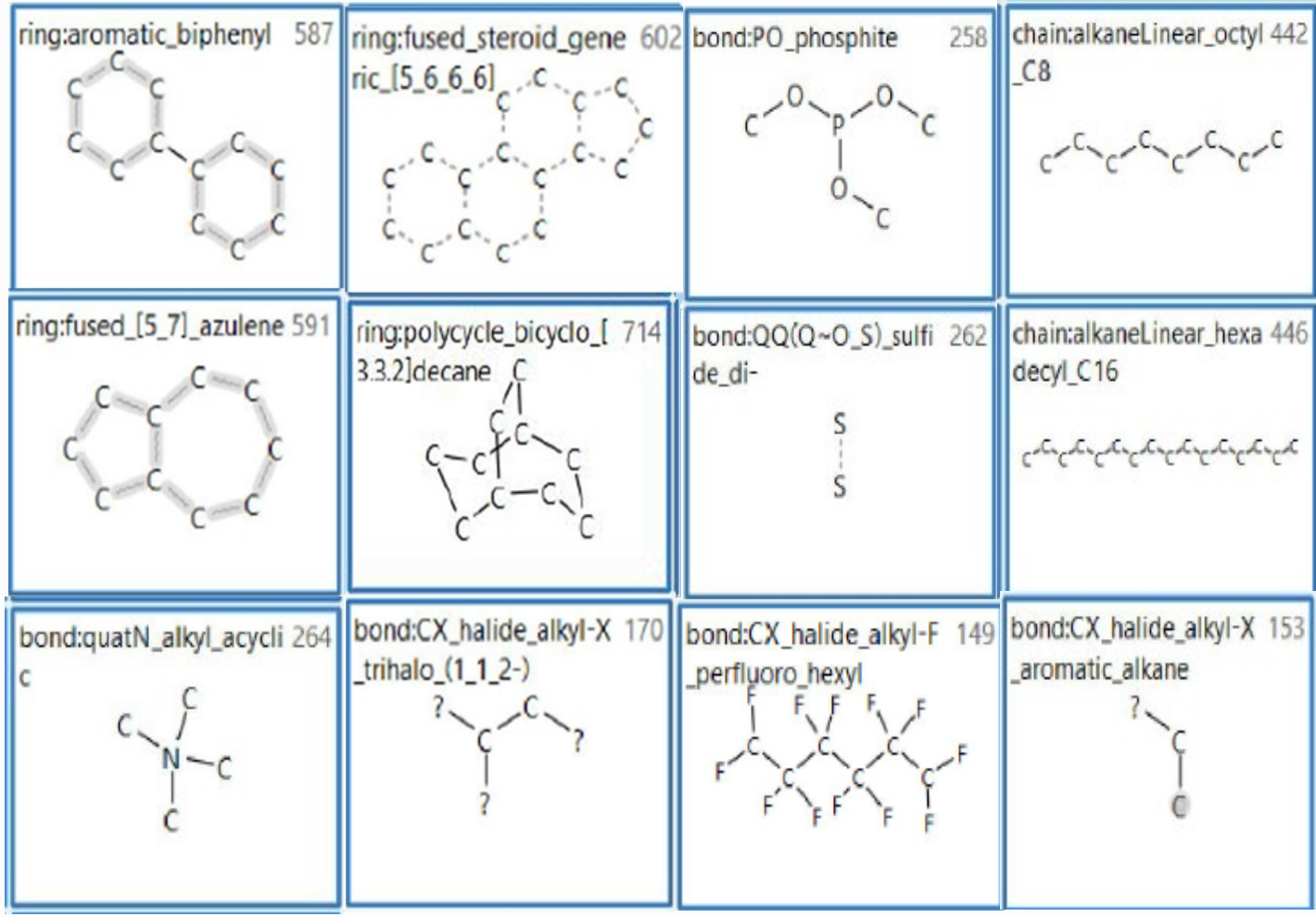
- Additional chemical coverage across all classes, with a slight increase in Unclassified MOAs relative to N/S classes
- Possible implications for N/S predictions

Characterize training set chemicals: ToxPrints

- Pull in chemotype information for our chemicals via ToxPrints (TxPs)
  - Publicly available tool
  - EPA Comptox Chemicals Dashboard

ToxPrints:

- ✓ *729 chemical features*
- ✓ *Chemically interpretable*
- ✓ *Coverage of diverse chemistry*
- ✓ *Hierarchical: Includes scaffolds, functional groups, chains, rings, bonding patterns, atom-types*



Yang et al. J. Chem. Inf. Model. 2015. Richard et al., Chem. Res. Toxicol. 2016, 29(8) 1225 – 1251; Strickland et al., Arch Toxicol. 2018 92(1) 487 – 500; Wang et al., Environment International 2019, 126 377 – 386

Classification model development

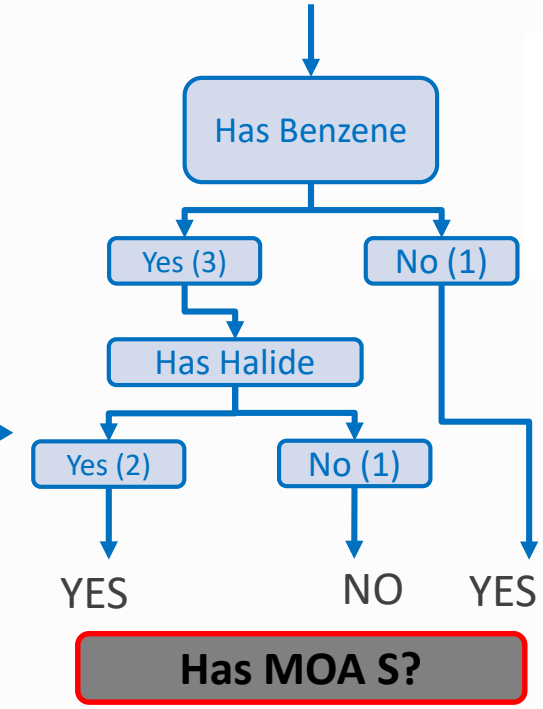
Figure adapted from Katherine Phillips

Train the Model

ToxPrints, NAM data or both

	cMOA	Has Metal	Has Halide	Has Benzene	Has Alkyne
Chemical 1	N	0	1	0	0
Chemical 2	N	0	0	1	1
Chemical 3	S	0	1	1	0
Chemical 4	S	1	1	1	1

cMOA Features



Repeated many times with different samples to build "forest" of classifier trees.

Valid models **must**:

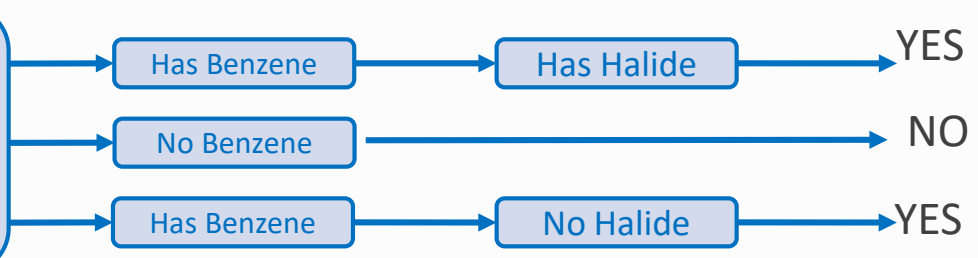
- accurately predict the training set
- predict beyond the training set
- be more predictive than a model built on randomized data

Predict with the Model

Chemical 5	0	1	1	0
Chemical 6	0	1	0	0
Chemical 7	0	0	1	1

Features

In Applicability Domain of Model?



Predict target with ***valid*** models using features

Repeated with each "tree"

p

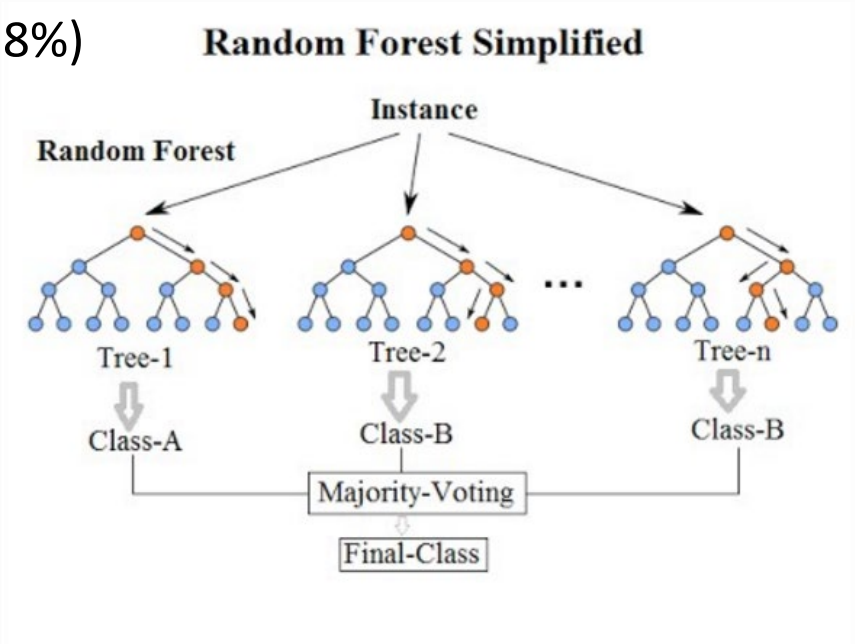
Probability for N or S

Office of Research and Development
Center for Computational Toxicology and Exposure

Classification model details

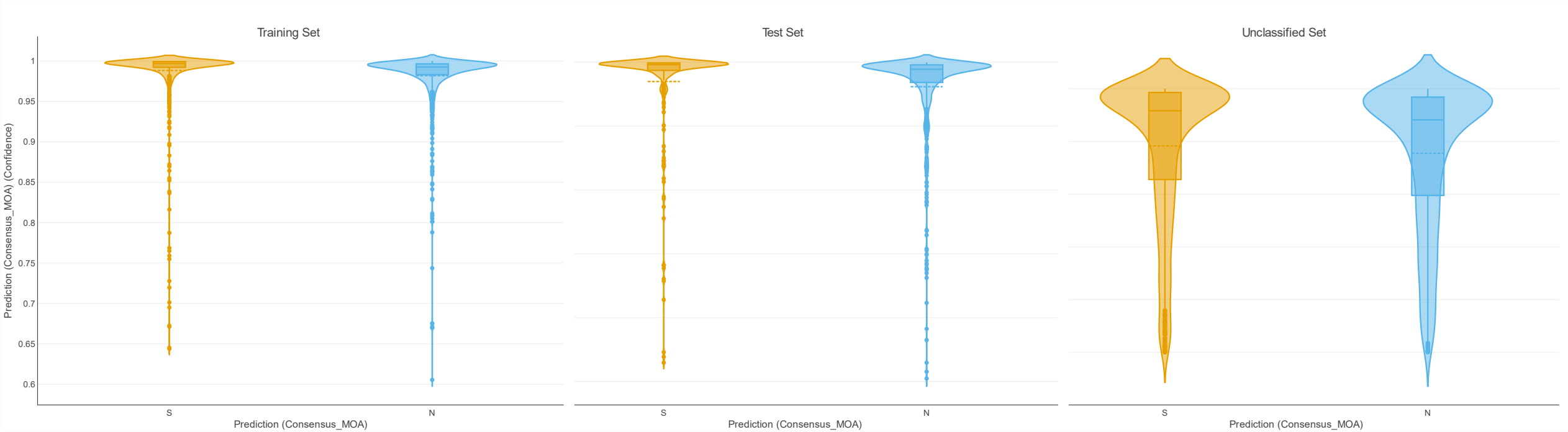
- Random Forest (Boosted Gradient Method) provided the best model results:
  - Split data into 80% training and 20% hold out (test) sets
  - Hyperparameter tuning with 5-fold cross validation, square-root sampling, etc.

  - Training set: “balanced” down-sampled subset (2104 chemicals w/ a cMOA = N or S)
  - High accuracy in both training and test sets (training = 99.7%; test = 95.8%)
  - Total Accuracy on all N + S data set = 97.6% (4356 cMOA = N or S)
    - Across all N + S chemicals -> 105 chemicals misclassified:
      - 24 F_{pos}{predicted S}
      - 81 F_{neg}{predicted N}



https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Distribution of prediction confidence [0,1] by (N,S) class



Training Set
Median: 0.999, 0.993
Mean: 0.988, 0.982

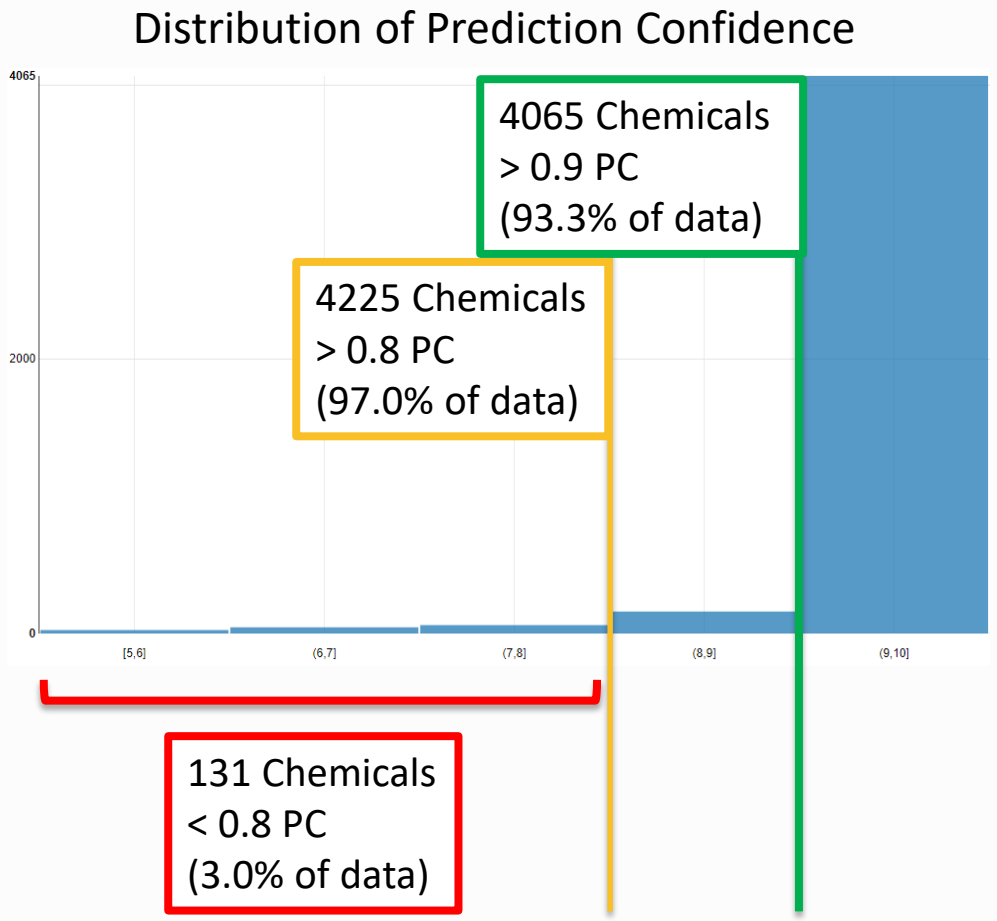
Test Set
Median: 0.996, 0.989
Mean: 0.970, 0.962

Unclassified Set
Median: 0.958, 0.941
Mean: 0.892, 0.877

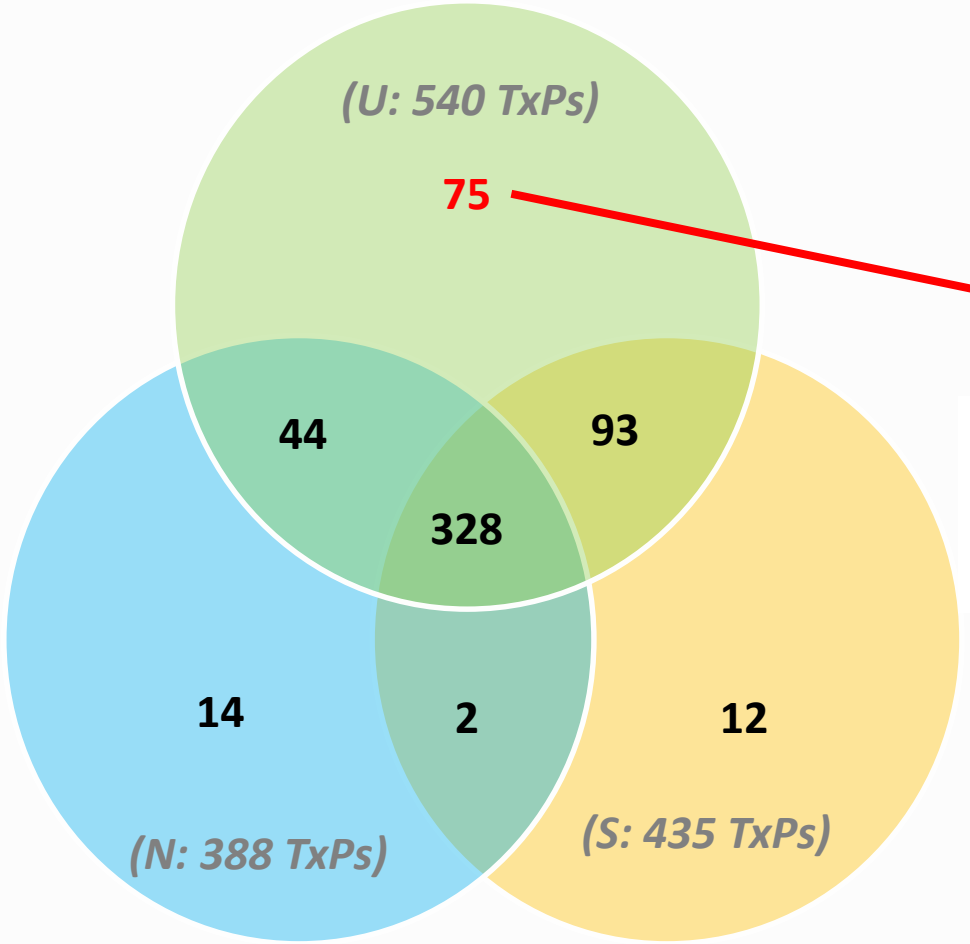
Specific-acting Narcotic

Prediction confidence across the cMOA = N or S

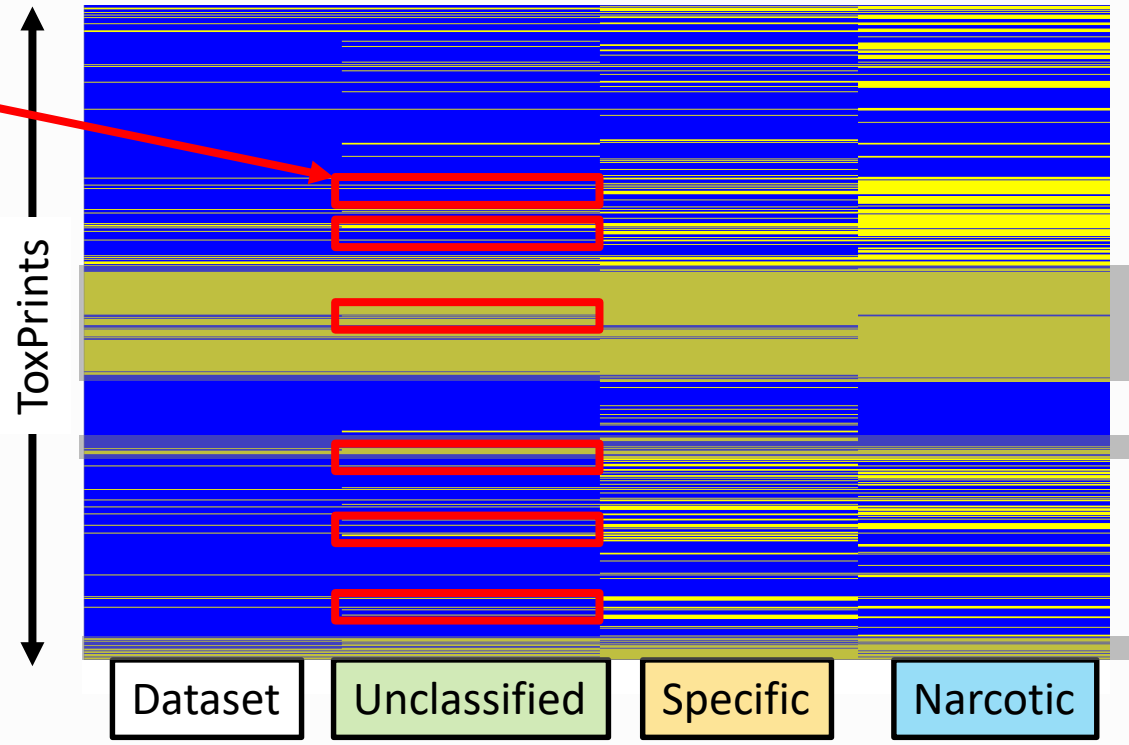
- Distribution of prediction confidence (PC) tends to be > 0.8 for the classified data (cMOA = N or S)
- Model has fewer # misclassifications in S
  - Misclassifications for 93 cMOA confidence = 2, and 12 with 1,3 scores (recall 3>2>1 for confidence)
  - ~46% of the misclassifications can be attributed to the chemicals with PC < 0.8
  - ~67% of the misclassification can be attributed to chemicals with PC < 0.88



Characterization of ToxPrint coverage across different classes



Heatmap representation of ToxPrints



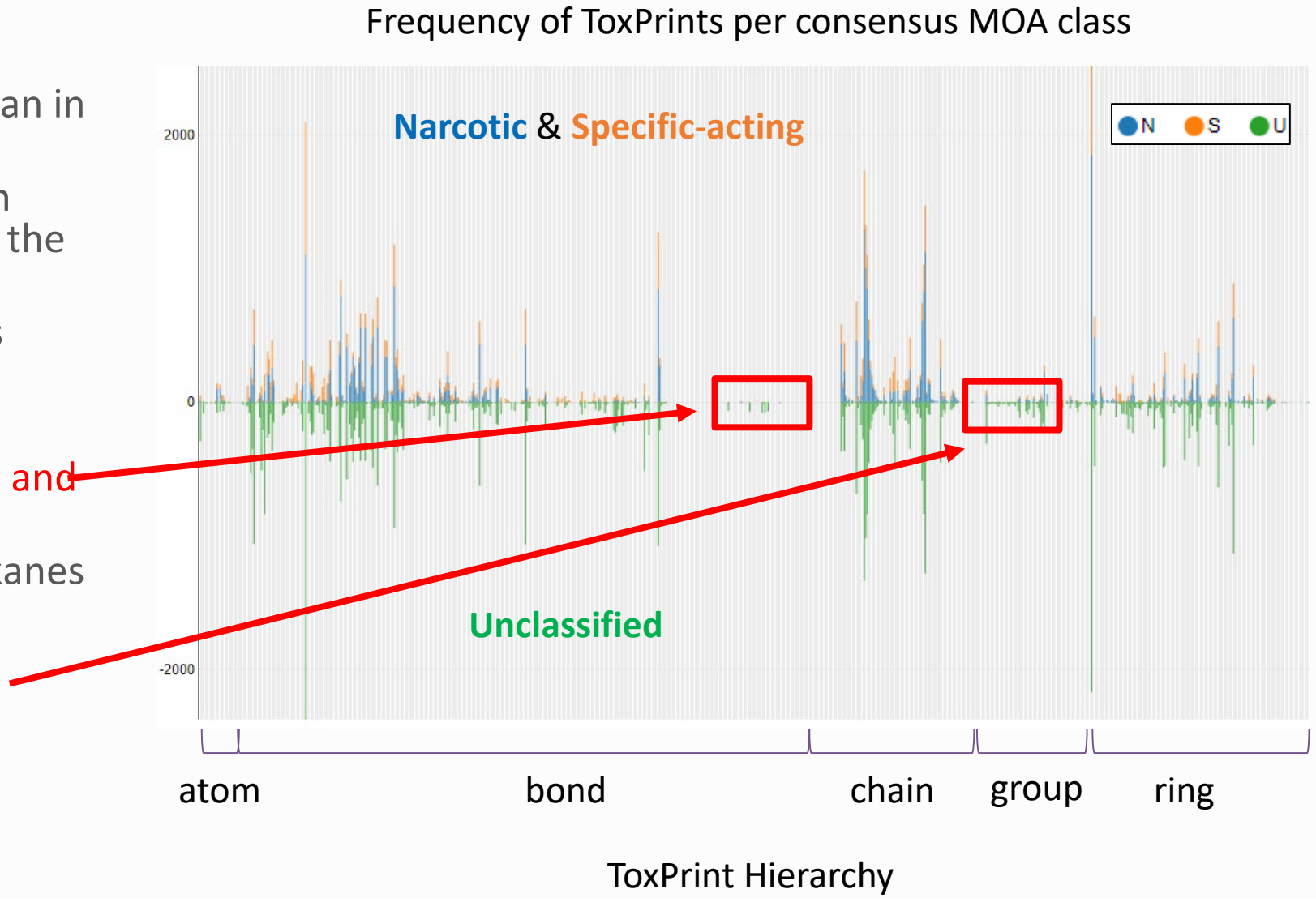
Missing structural Domains across full dataset:

- Metal & metalloid bonds
- Propyleneoxide chains
- Bicyclo ring structures

ToxPrints: Dataset > Unclassified > Specific-acting > Narcotic

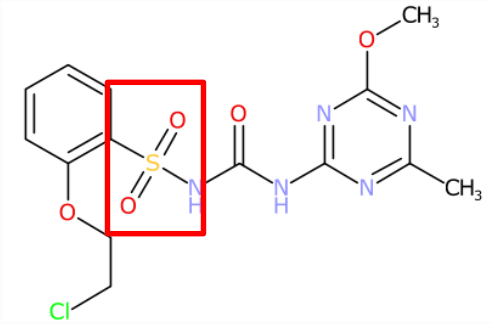
What are these 75 unique ToxPrints in the Unclassified set?

- ~7x more unique features in U (than in N or S)
- Could explain the lower prediction confidence in N/S classification of the U set
- Potential for additional categories based on structures:
  - 2 atom TxPs (metal group III)
  - 38 bond TxPs (metalloid: silane and siloxanes…)
  - 8 chain TxPs (ethyleneoxide alkanes C10 – C20)
  - 19 group TxPs (amino acids, polydentate ligands)
  - 8 ring TxPs

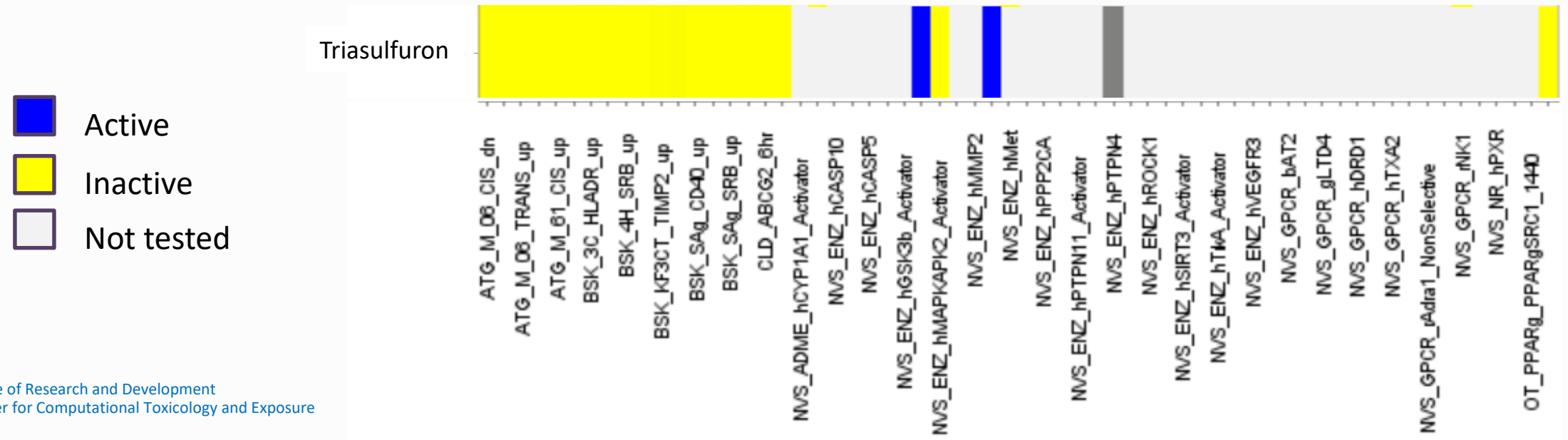


Example: Differences in model prediction vs. cMOA: Triasulfuron

- N-sulfonylurea herbicide
- Model prediction: Specific-acting
- EnviroTox consensus MOA: Narcotic
- ECOSAR classification: Sulfonyl Urea
- S(=O)_sulfonyl ToxPrint is enriched in the specific-acting MOA space and 47 assays



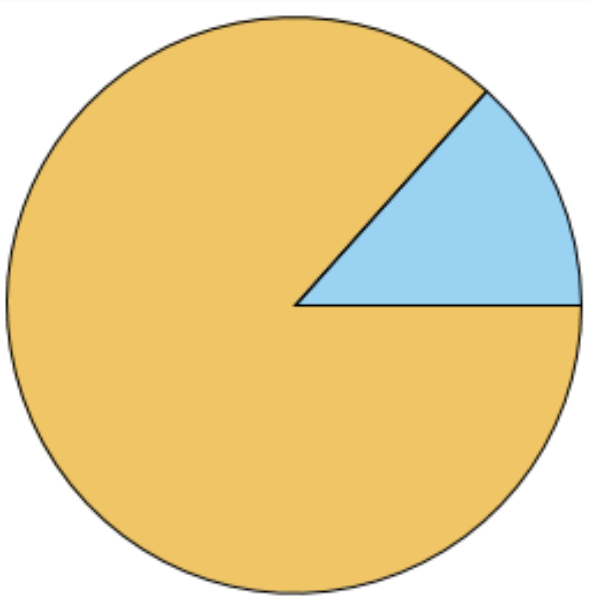
CASRN 82097-50-5
DTXSID0024345

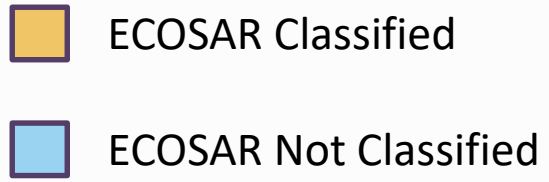


Preliminary predicted MOAs of the EnviroTox Unclassified set

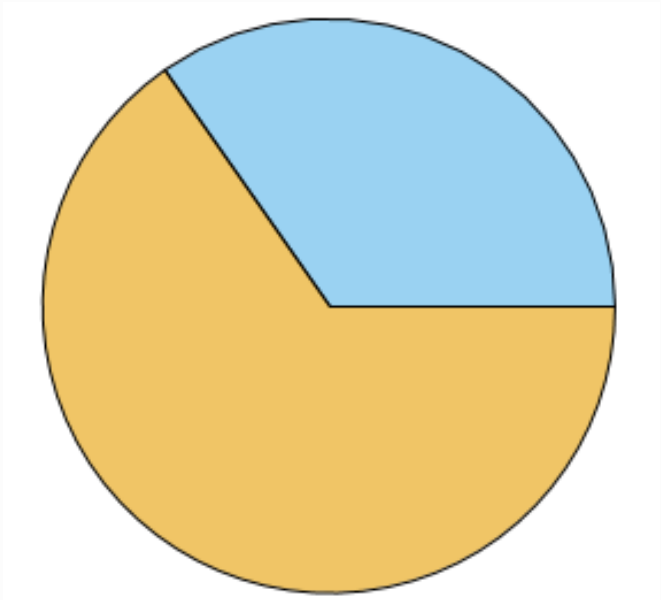
- 674 chemicals in the EnviroTox dataset that had low confidence or ambiguous consensus
- Applied model to the Unclassified set and compared predictions to ECOSAR classification
- ***Currently extending this analysis to the additional 3089 unclassified chemicals***

361 predicted as Narcotic





313 predicted as Specific-acting



Unclassified chemicals, predicted Specific-Acting: Enriched ToxPrints

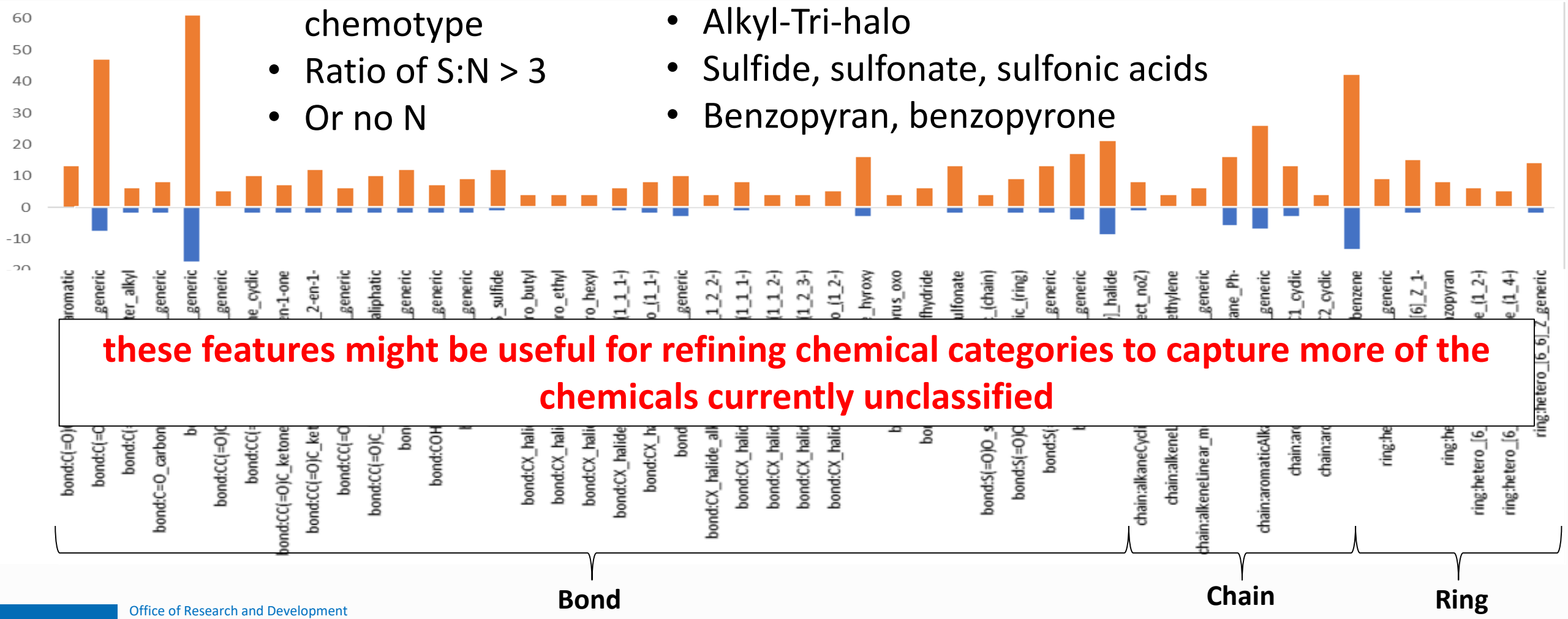
United States Environmental Protection Agency

Criteria:

- ≥ 3 chemicals per chemotype
- Ratio of S:N > 3
- Or no N

Results:

- Ketones
- Alkyl-Tri-halo
- Sulfide, sulfonate, sulfonic acids
- Benzopyran, benzopyrone



Summary

- Identified relevant NAM information to develop a classification model for specific-acting MOAs
  - Increased the available chemical space of EnviroTox
- Explored differences in predicted and consensus MOA via chemotype enrichments
- Used model to inform ECOSAR preliminary set of unclassified chemicals
  - Majority of unclassified chemicals were predicted to have a specific acting MOA
  - Identified primary chemotypes for specific acting MOAs

- Use methods to inform classification models for TSCA (New Chemical Categories)
- Use chemotype enrichments to identify potential bioassays with bioactivity to provide support of NAM data in category development

United States Environmental Protection Agency

Thank you!