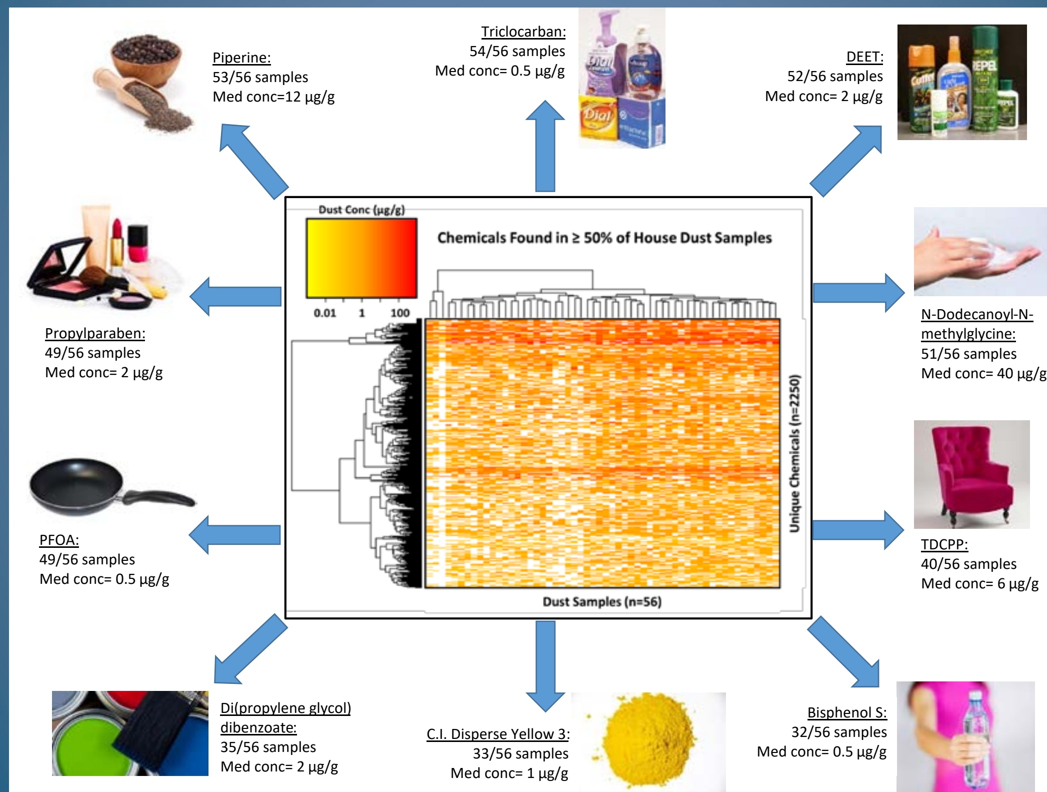
Advancing Non-Targeted Analysis Research within EPA/ORD



Jon R. Sobus
US EPA Office of Research and Development

July 28, 2016

Comparing Analysis Approaches

- Targeted Analysis:
  - We know exactly what we're looking for
  - 10s – 100s of chemicals
- Suspect Screening Analysis (SSA):
  - We have chemicals of interest
  - 100s – 1,000s of chemicals
- Non-Targeted Analysis (NTA):
  - We have no preconceived notions or lists
  - 1,000s – 10,000s of chemicals
    - In dust, soil, food, air, water, products, plants, animals, and us!!

High Throughput Screening Methods

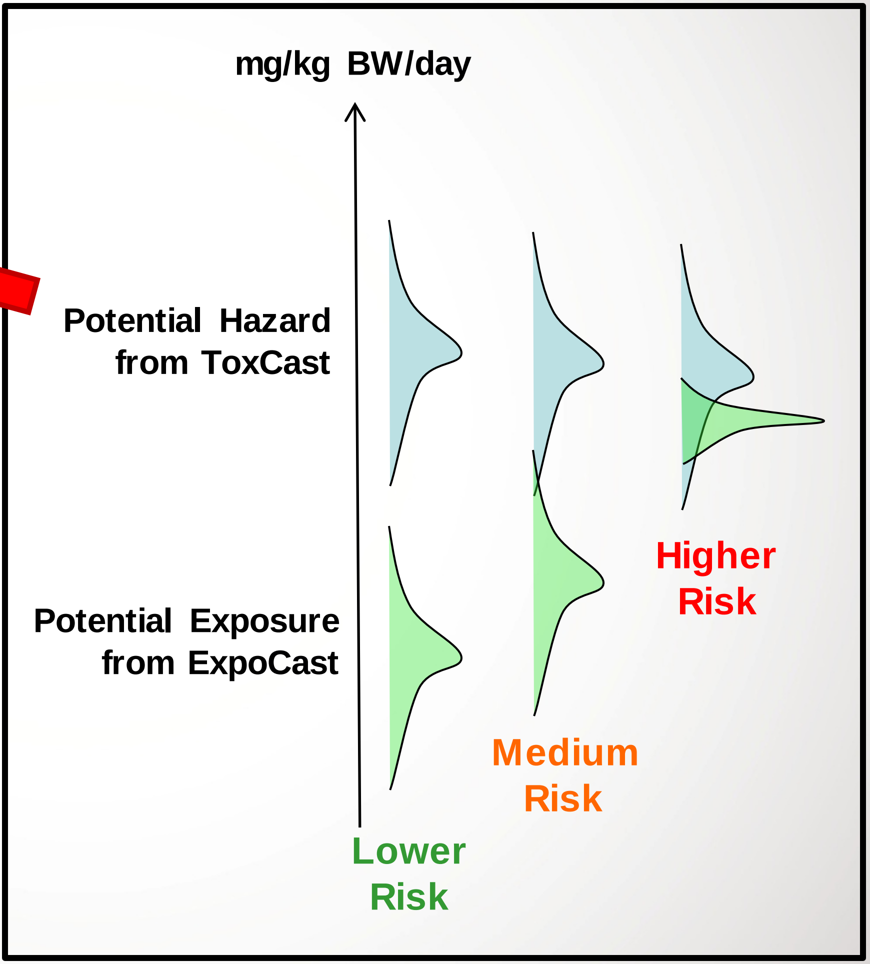
Research and Testing Needs

Nominations for:

1. Parent chemicals
2. Mixtures
3. Metabolites/Degradates

Measurement data for:

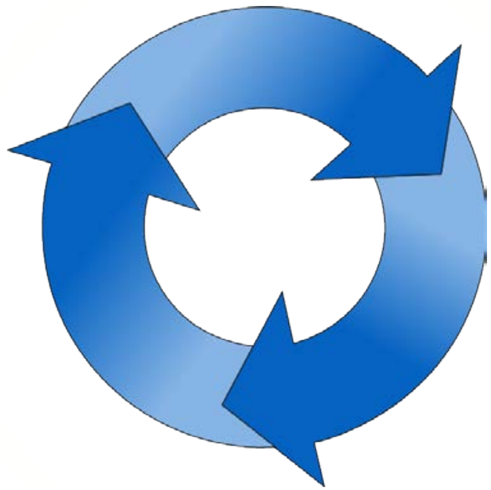
1. Model inputs
2. Model evaluation
3. Model refinement



Currently ~8000 chemicals

Tools of the Trade

Analytical Instruments



Comp. Tools & Workflows

FOR IDENT

XCMS Online

MetFrag

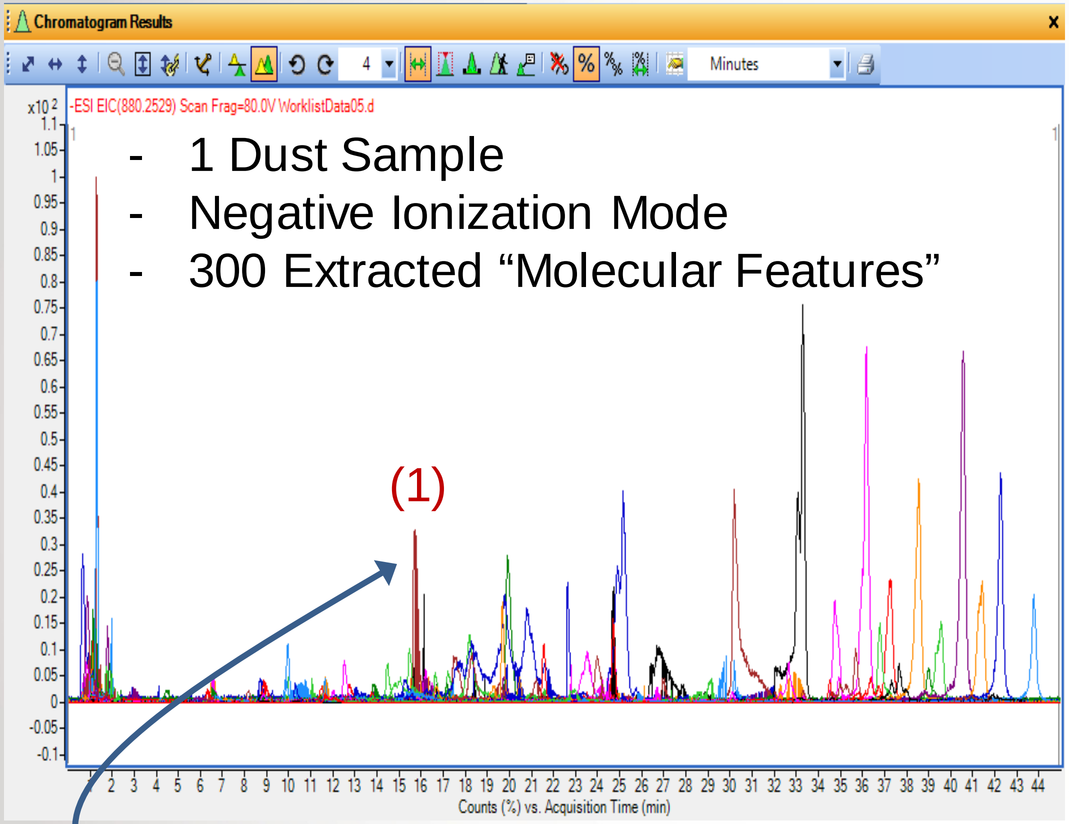
Chemical Databases

ChemSpider
The free chemical database

PubChem

MassBank

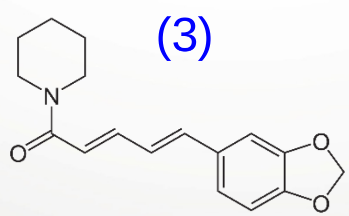
General Goals of SSA/NTA



1) Prioritize "Molecular Features"
2) Correctly assign formulas
3) Correctly assign structures
4) Determine chemical sources
5) Predict chemical concentrations

(2)

$C_{17}H_{19}NO_3$ →



(3)

→

(4)

→

(5)

12 µg/g

Previous Work with SSA

Environment International 88 (2016) 269–280

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

ELSEVIER

Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring

CrossMark

Julia E. Rager [a], Mark J. Strynar [b], Shuang Liang [a], Rebecca L. McMahen [a], Ann M. Richard [c], Christopher M. Grulke [d], John F. Wambaugh [c], Kristin K. Isaacs [b], Richard Judson [c], Antony J. Williams [c], Jon R. Sobus [b,*]

[a] Oak Ridge Institute for Science and Education (ORISE) Participant, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[b] U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[c] U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States
[d] Lockheed Martin, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, United States

SSA Workflow

Dust samples (<150 um) (n=56) Taken from National Survey

Extract and Analyze Samples (LC-TOF/MS)

↓

Identify Molecular Features (User-defined criteria)

↓

Assign Formulas to Molecular Features (DSSTox-MSMF Database)

↓

Link Assigned Formulas to Chemicals/Structures (DSSTox_v2 Database)

↓

- Estimate Average **Abundance (A)** and **Number of Samples (N)** Associated with Each Chemical
- Group Chemicals Into **Exposure (E)** Categories Using ExpoCast
- Assess Chemical **Bioactivity (B)** Using Tox21

↓

- Group A: For Chemicals with **E** and **B**, Prioritization Score = f(**A**+**N**+**E**+**B**)
- Group B: For Chemicals without **E** and **B**, Prioritization Score = f(**A**+**N**)

↓

Confirm Chemicals with High Priority Scores Using Standards

Agilent Technologies

EPA DSSTox Structure-Browser GO v 2.0

Molecular Features in Dust

~3000 features identified per sample

Number of features identified varied between samples

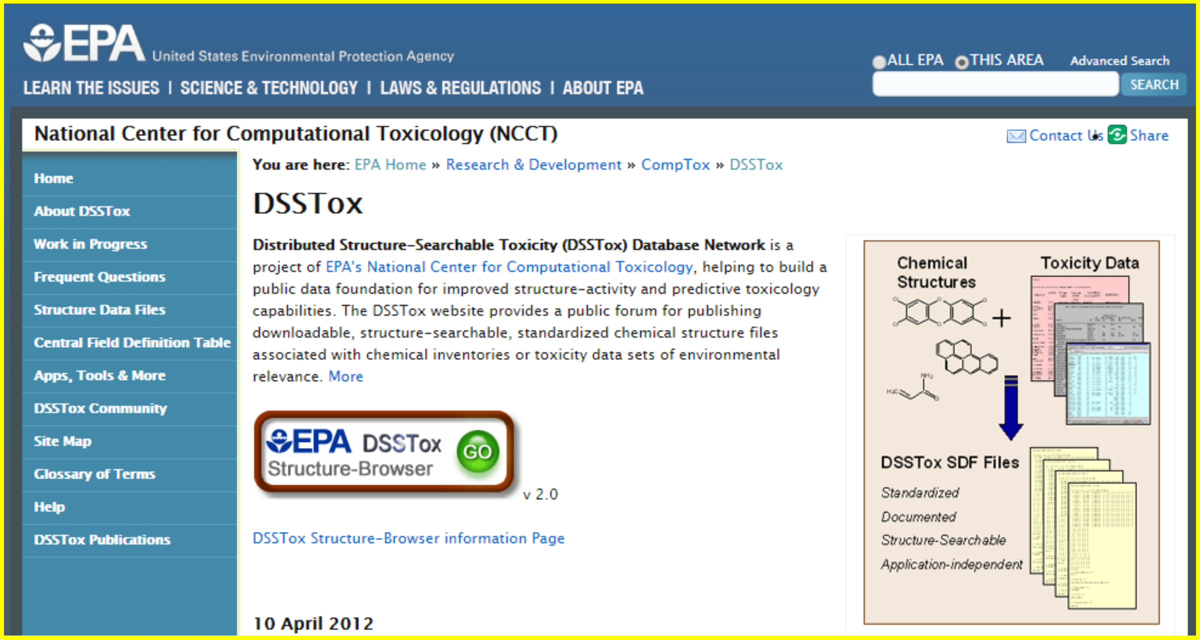
- 10-fold range (max/min) in positive mode
- 15-fold range (max/min) in negative mode

Positive Ionization Mode					
	Mean	SD	Min	Med	Max
Abundance	$9.32x10^5$	$3.94x10^6$	$1.46x10^4$	$2.61x10^5$	$2.33x10^8$
Number of Features per Sample	3185	1023	632	3262	5477
Number of Formula Matches per Sample	45	14	4	45	77
Negative Ionization Mode					
	Mean	SD	Min	Med	Max
Abundance	$1.26x10^6$	$7.87x10^6$	$1.61x10^4$	$2.58x10^5$	$6.06x10^8$
Number of Features per Sample	2236	646	260	2169	3739
Number of Formula Matches per Sample	44	27	10	38	116

Chemical Database (DSSTox)

- Carefully curated database
- Standardized chemical mass, formula, structure
- One-to-one mapping of CAS-to-chemical name
- Environmental contaminants, pharmaceuticals, industrial chemicals, etc.

•~33K chemicals in DSSTox <u>at time of dust SSA analysis</u>



Formulas Identified in Dust

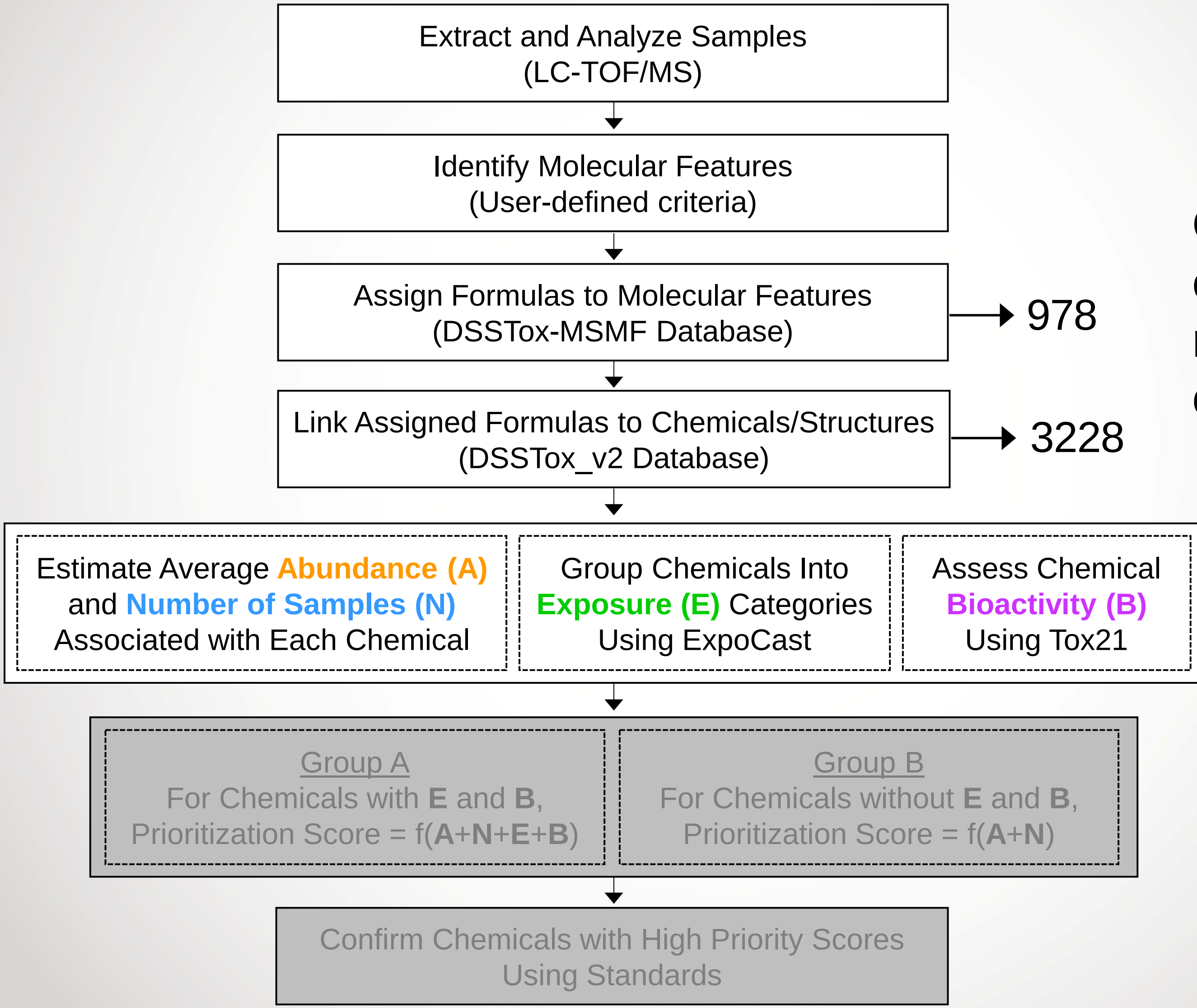
Required strict match score of ≥ 90

~45 formulas tentatively identified per sample, per mode, on average

Represents < 2% of the total # of observed features

Positive Ionization Mode					
	Mean	SD	Min	Med	Max
Abundance	9.32×10^5	3.94×10^6	1.46×10^4	2.61×10^5	2.33×10^8
Number of Features per Sample	3185	1023	632	3262	5477
Number of Formula Matches per Sample	45	14	4	45	77
Negative Ionization Mode					
	Mean	SD	Min	Med	Max
Abundance	1.26×10^6	7.87×10^6	1.61×10^4	2.58×10^5	6.06×10^8
Number of Features per Sample	2236	646	260	2169	3739
Number of Formula Matches per Sample	44	27	10	38	116

SSA Workflow



On average every formula represents 3 chemicals

Exposure Estimates from ExpoCast

- 5 exposure descriptors used to estimate exposure to ~8000 chemicals

- Exposure rates grouped into categories (based on estimated median values for U.S. population):

Category 1 < 1×10^{-8} mg/kg/day;
Category 2 > 1×10^{-8} and < 1×10^{-7} mg/kg/day;
Category 3 > 1×10^{-7} and < 1×10^{-6} mg/kg/day;
Category 4 > 1×10^{-6} and < 1×10^{-5} mg/kg/day;
Category 5 > 1×10^{-5} and < 1×10^{-4} mg/kg/day;
Category 6 > 1×10^{-4} and < 1×10^{-3} mg/kg/day;
Category 7 > 1×10^{-3} and < 1×10^{-2} mg/kg/day

Environmental Science & Technology
Article
pubs.acs.org/est

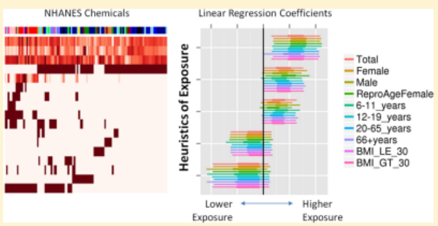
High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals

John F. Wambaugh,[*,†] Anran Wang,[†,§,∥] Kathie L. Dionisio,[‡] Alicia Frame,[†,∥] Peter Egeghy,[‡] Richard Judson,[†] and R. Woodrow Setzer[†]

[†]National Center for Computational Toxicology, and [‡]National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States
[§]North Carolina State University, Department of Statistics, Raleigh, North Carolina 27695-8203, United States
[∥]Oak Ridge Institute for Science and Education Grantee, P.O. Box 117, Oak Ridge, Tennessee 37831-0117, United States

Supporting Information

ABSTRACT: The risk posed to human health by any of the thousands of untested anthropogenic chemicals in our environment is a function of both the hazard presented by the chemical and the extent of exposure. However, many chemicals lack estimates of exposure intake, limiting the understanding of health risks. We aim to develop a rapid heuristic method to determine potential human exposure to chemicals for application to the thousands of chemicals with little or no exposure data. We used Bayesian methodology to infer ranges of exposure consistent with biomarkers identified in urine samples from the U.S. population by the National Health and Nutrition Examination Survey (NHANES). We performed linear regression on inferred exposure for demographic subsets of NHANES demarked by age, gender, and weight using chemical descriptors and use information from multiple databases and structure-based calculators. Five descriptors are capable of explaining roughly 50% of the variability in geometric means across 106 NHANES chemicals for all the demographic groups, including children aged 6−11. We use these descriptors to estimate human exposure to 7968 chemicals, the majority of which have no other quantitative exposure prediction. For thousands of chemicals with no other information, this approach allows forecasting of average exposure intake of environmental chemicals.



Bioactivity Data from Tox21

High-throughput toxicity screening data on >8,000 chemicals

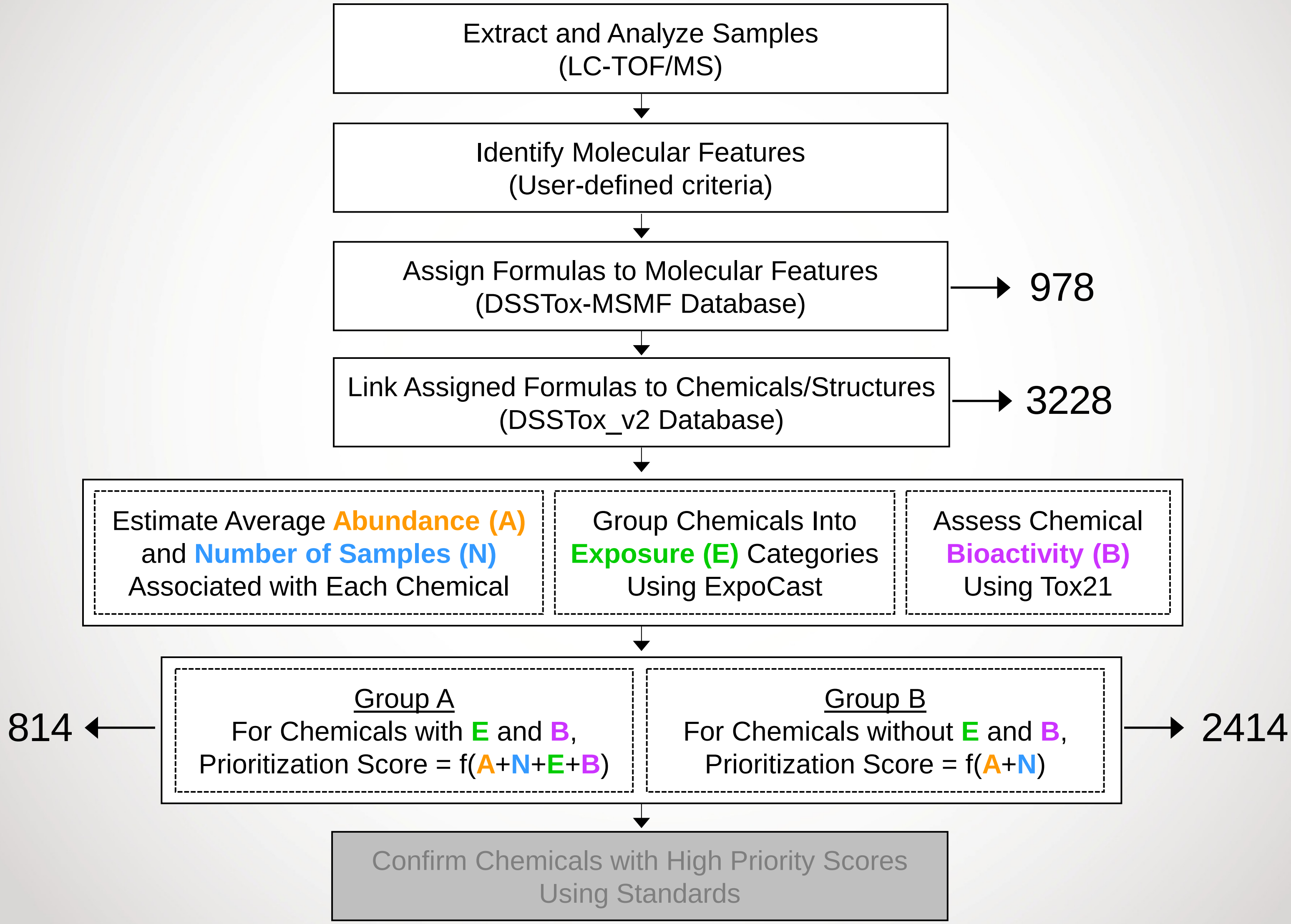
http://www.epa.gov/ncct/Tox21/

Tox21 data used here:

Hit calls (0=inactive, 1=active) for:

- AhR (aryl hydrocarbon receptor)
- AR (androgen receptor)
- ERα (estrogen receptor 1)
- NFκB1 (nuclear factor of kappa light polypeptide gene enhancer in B cells 1)
- PPARγ (peroxisome proliferator-activated receptor gamma)

SSA Workflow

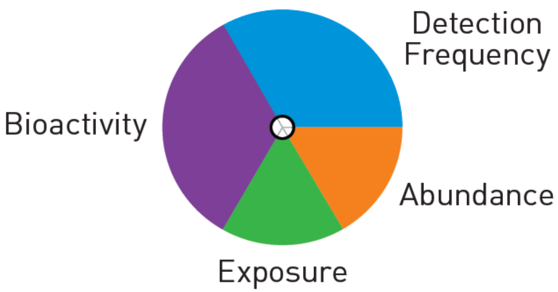


Prioritization Scoring with ToxPi

$$\text{ToxPi Score}_i = w_A \frac{A_i - A_{min}}{A_{max} - A_{min}} + w_N \frac{N_i - N_{min}}{N_{max} - N_{min}} + w_E \frac{E_i - E_{min}}{E_{max} - E_{min}} + w_B \frac{B_i - B_{min}}{B_{max} - B_{min}}$$

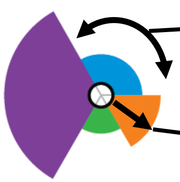
$$w_A = w_E = 1; \quad w_N = w_B = 2$$

ToxPi Legend



Individual components of a unit circle are scaled and represented as "slices"

Example Chemical

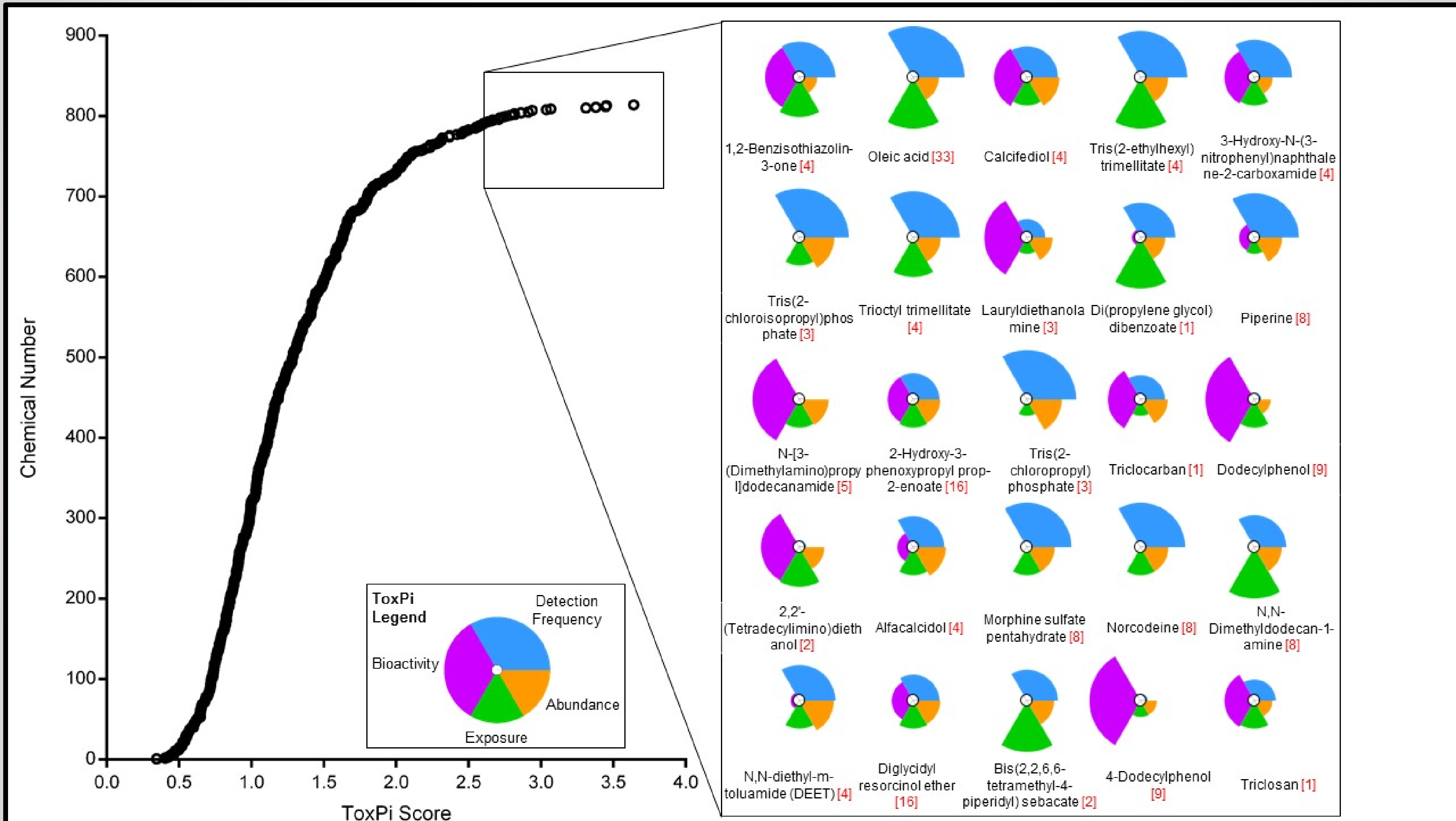


Width indicates the relative weight of the variable

Distance from the origin is proportional to the normalized value of the data

(Reif et al. 2010)

Group A Priority Candidates*



*listed chemicals are not necessarily confirmed

SSA Workflow

Extract and Analyze Samples
(LC-TOF/MS)

↓

Identify Molecular Features
(User-defined criteria)

↓

Assign Formulas to Molecular Features
(DSSTox-MSMF Database)

↓

Link Assigned Formulas to Chemicals/Structures
(DSSTox_v2 Database)

↓

Estimate Average **Abundance (A)** and **Number of Samples (N)** Associated with Each Chemical	Group Chemicals Into **Exposure (E)** Categories Using ExpoCast	Assess Chemical **Bioactivity (B)** Using Tox21

↓

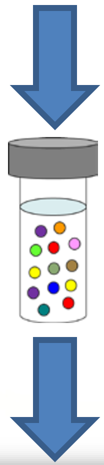
Group A	Group B
For Chemicals with **E** and **B**, Prioritization Score = f(**A**+**N**+**E**+**B**)	For Chemicals without **E** and **B**, Prioritization Score = f(**A**+**N**)

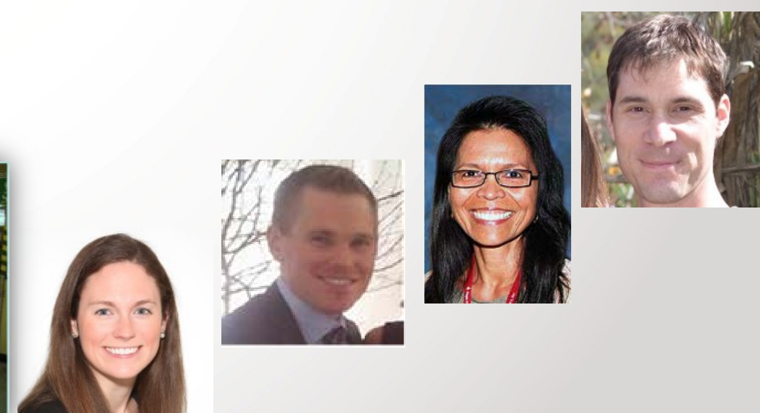
↓

Confirm Chemicals with High Priority Scores
Using Standards

Blinded Analysis of 100-Chemical Mixture

ToxCast™





Blinded Analysis: Procedures & Results

- Analyzed at 2 µM and 0.2 µM, neg. and pos. modes
- Logical scheme used to rank features from 0 to 5 stars
  - Present at both concentrations (>3x difference in response)
  - Consistent retention times
  - Match score ≥ 90
  - Peak saturation?
- Matching to dust features using formula, RT & spectra

100 Total Chemicals

→ 70 Detected Across Both Modes

→ 51 of Minimally-Sufficient Quality

→ 33 Matches in House Dust

Results for Chemicals Confirmed in House Dust

Chemical Name	ToxPi Rank (%)	N_{true}	SciFinder hits
Di(propylene glycol) dibenzoate	1.1	4	0
Piperine	1.2	42	1
Triclocarban	1.7	21	0
N,N-diethyl-m-toluamide (DEET)	2.6	33	22
Diethyl phthalate (DEP)	4.2	23	36
Propylparaben	5.4	19	7
3,6,9,12-Tetraoxahexadecan-1-ol	5.7	1	0
N-Dodecanoyl-N-methylglycine	6.0	6	0
Tris(1,3-dichloro-2-propyl) phosphate (TDCPP)	6.8	15	38
Methylparaben	8.7	16	10
Carbamazepine	12.0	1	0
Tris(2-ethylhexyl) phosphate (TEHP)	12.4	1	18
2-[2-(2-Butoxyethoxy)ethoxy]ethanol	15.5	2	2
Triethyl citrate	16.8	6	0
Tetradecanoic acid, 2,3-dihydroxypropyl ester	18.3	1	0
Clorophene	25.1	4	0
Nicotine	25.3	10	24
4,4'-Sulfonyldiphenol	33.5	4	1
Perfluoroctylsulfonamide acid (PFOSA)	34.4	1	9
Fluconazole	34.8	1	0
Perfluorooctanoic acid (PFOA)	38.0	3	33
Corticosterone	39.9	1	3
Dibutyl hexanedioate	48.9	1	3
Phosphoric acid, dibutyl ester	51.0	4	1
C.I. Disperse Yellow 3	51.4	3	0
Octyl beta-D-glucopyranoside	51.7	1	0
Perfluorodecanoic acid (PFDA)	54.2	3	13
Carbaryl	55.5	2	15
Rofecoxib	77.1	1	0
Primidone	78.6	3	0
2,4,5-Trichlorobenzenesulfonic acid	82.7	2	0
Lufenuron	89.7	1	0
Diphenyl phosphate	91.4	6	3

45% of confirmed chemicals not previously studied in house dust?

We’re on the Right Path…

- … but certainly room for improvement
- ~300,000 total molecular features (not unique)
- 33 confirmed chemicals
- State-of-the-art SSA yields <5% confirmed IDs
- So what else is in these (and other) samples??

Integrating SSA and NTA Workflows

Suspect Screening

- Raw Sample
- Extracted Sample
- Raw Features
- "Molecular Features"
- DSSTox Chemical Library
- Matched Formulas
- Mapped Structures
- Prioritized Structures (using ToxPi)
- Confirmed Structures (using ToxCast standards)

SSA workflow from Rager *et al.* analysis

Color Key

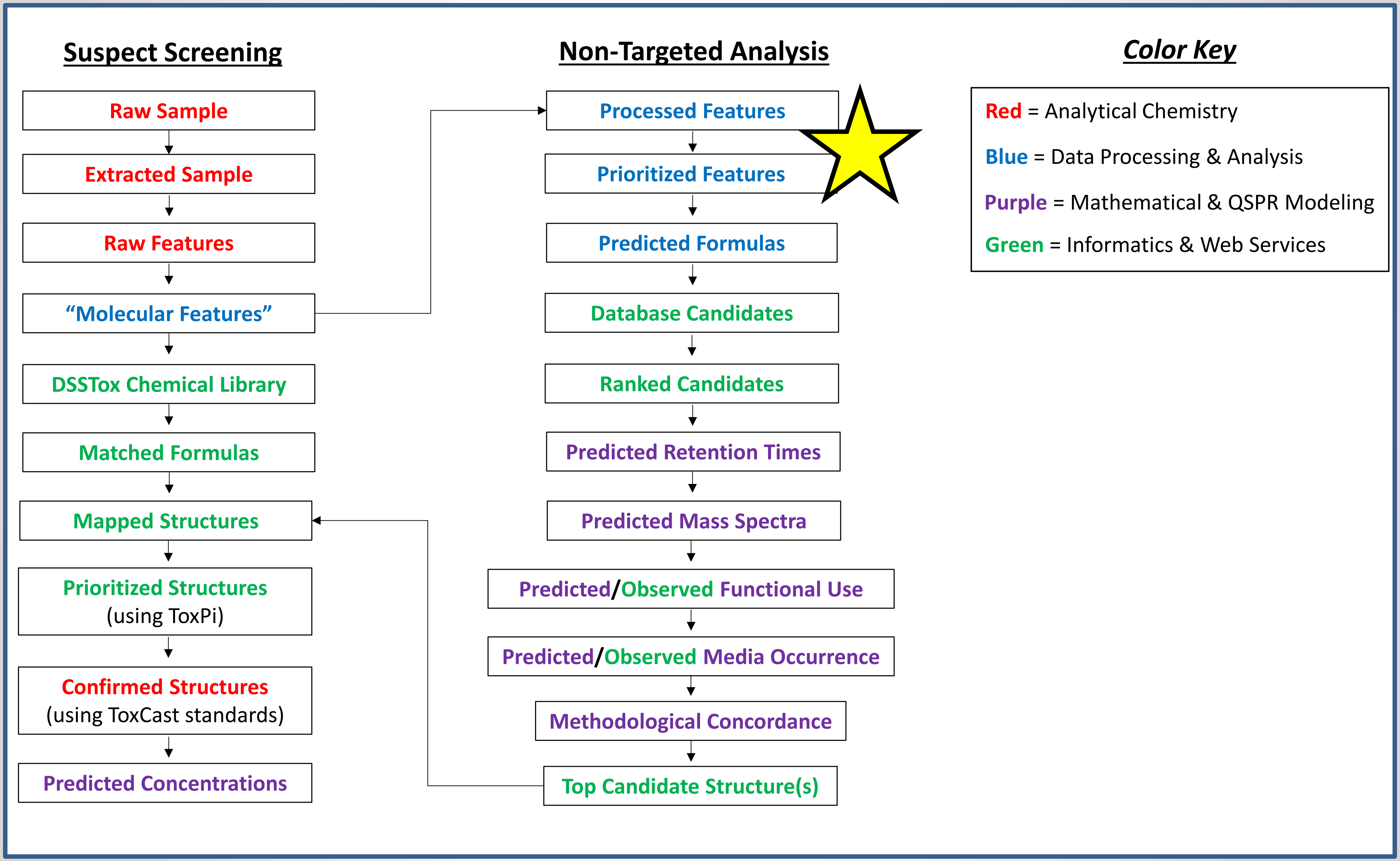
Red = Analytical Chemistry

Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

Green = Informatics & Web Services

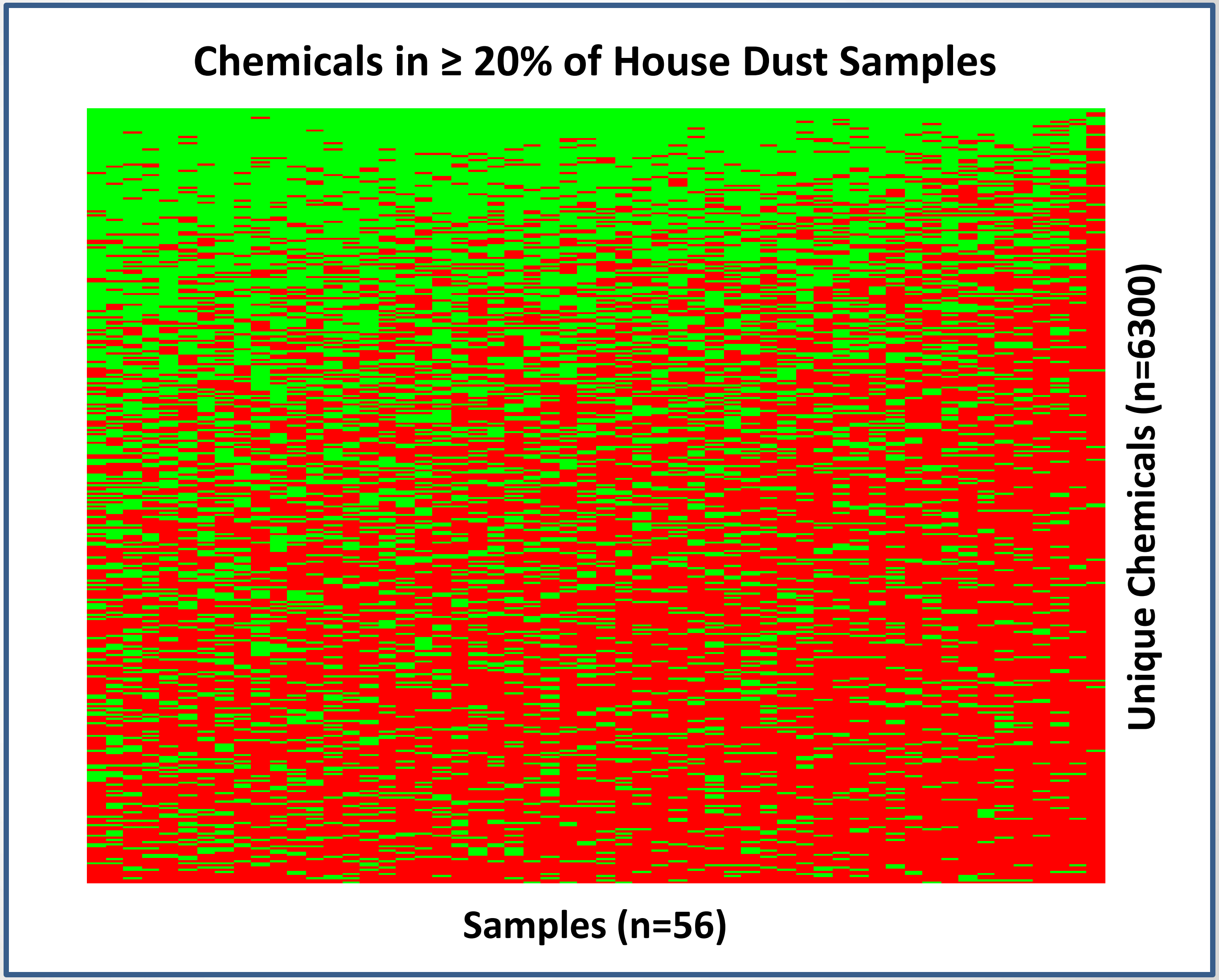
Feature Processing and Prioritization



Alignment of All Features Across Samples

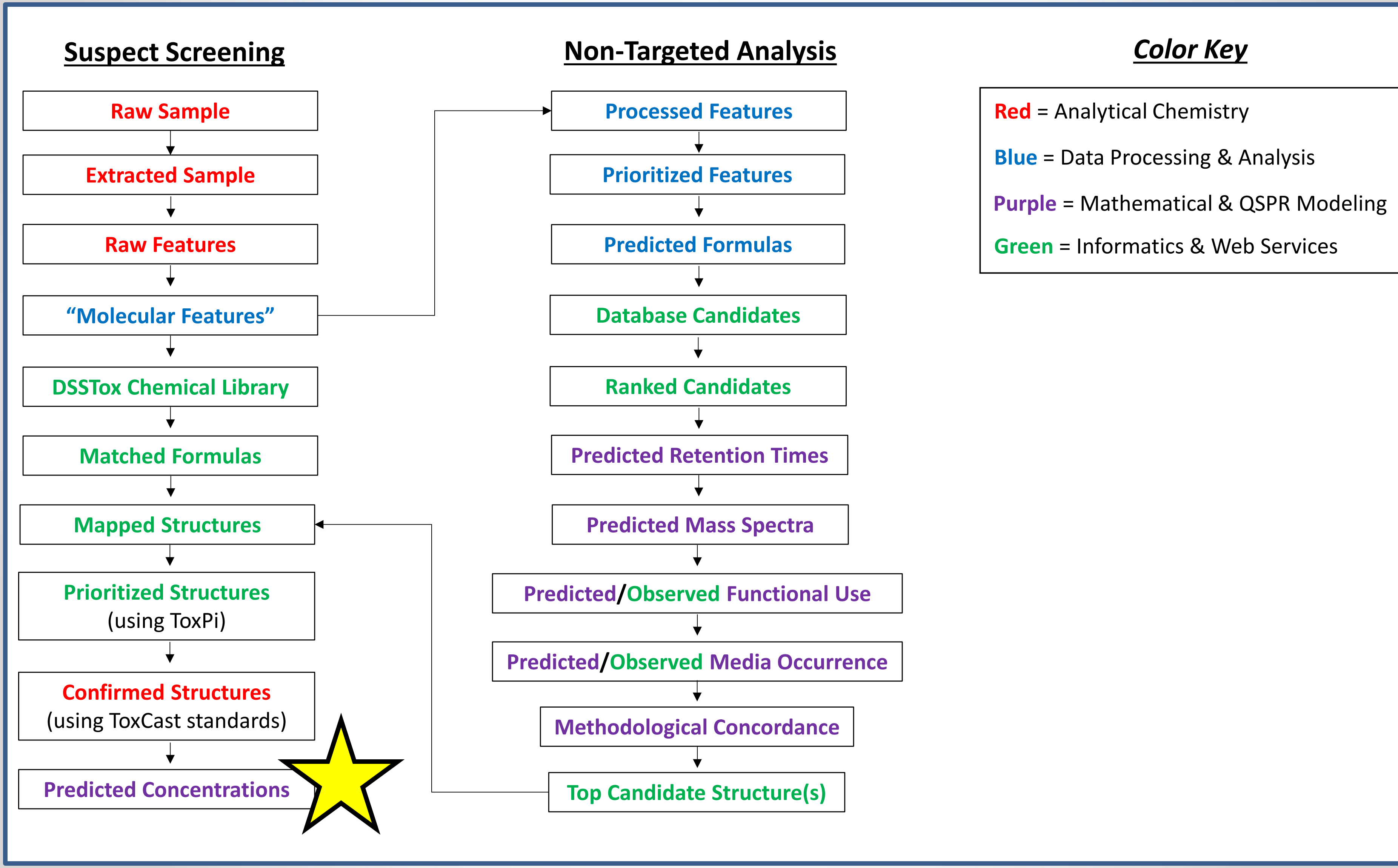
Most frequently occurring

Least frequently occurring

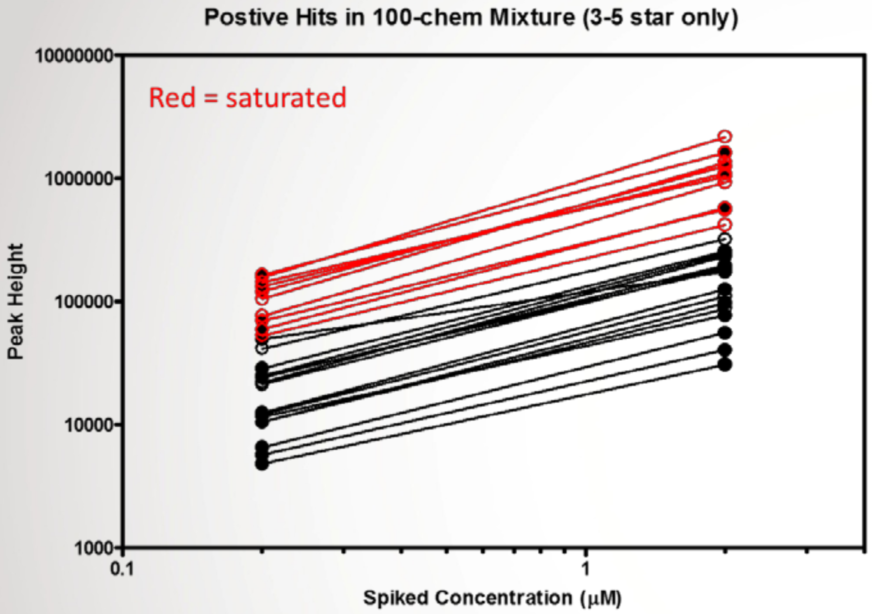


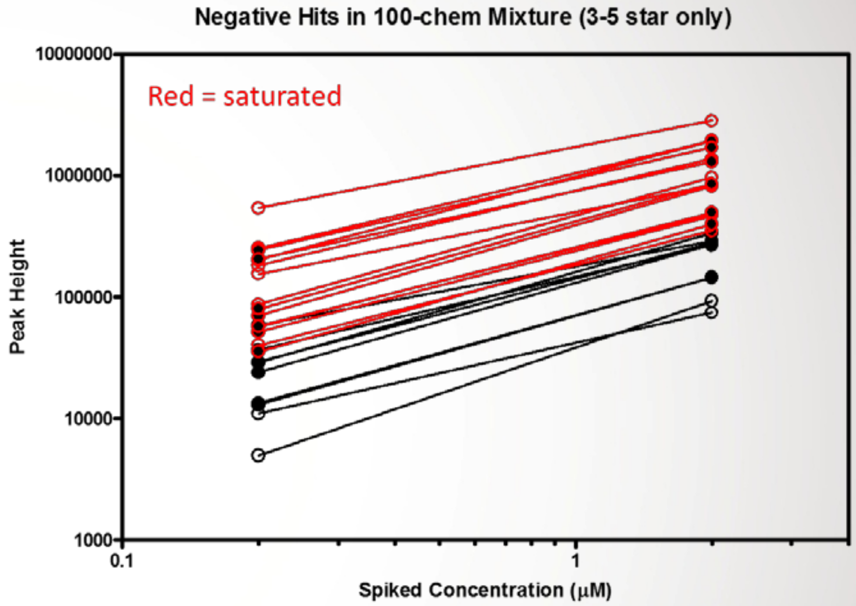
~ 80K total features across 56 samples

Estimating Medium-Specific Concentrations



Global Cal. Curves from 100-chem Mixture

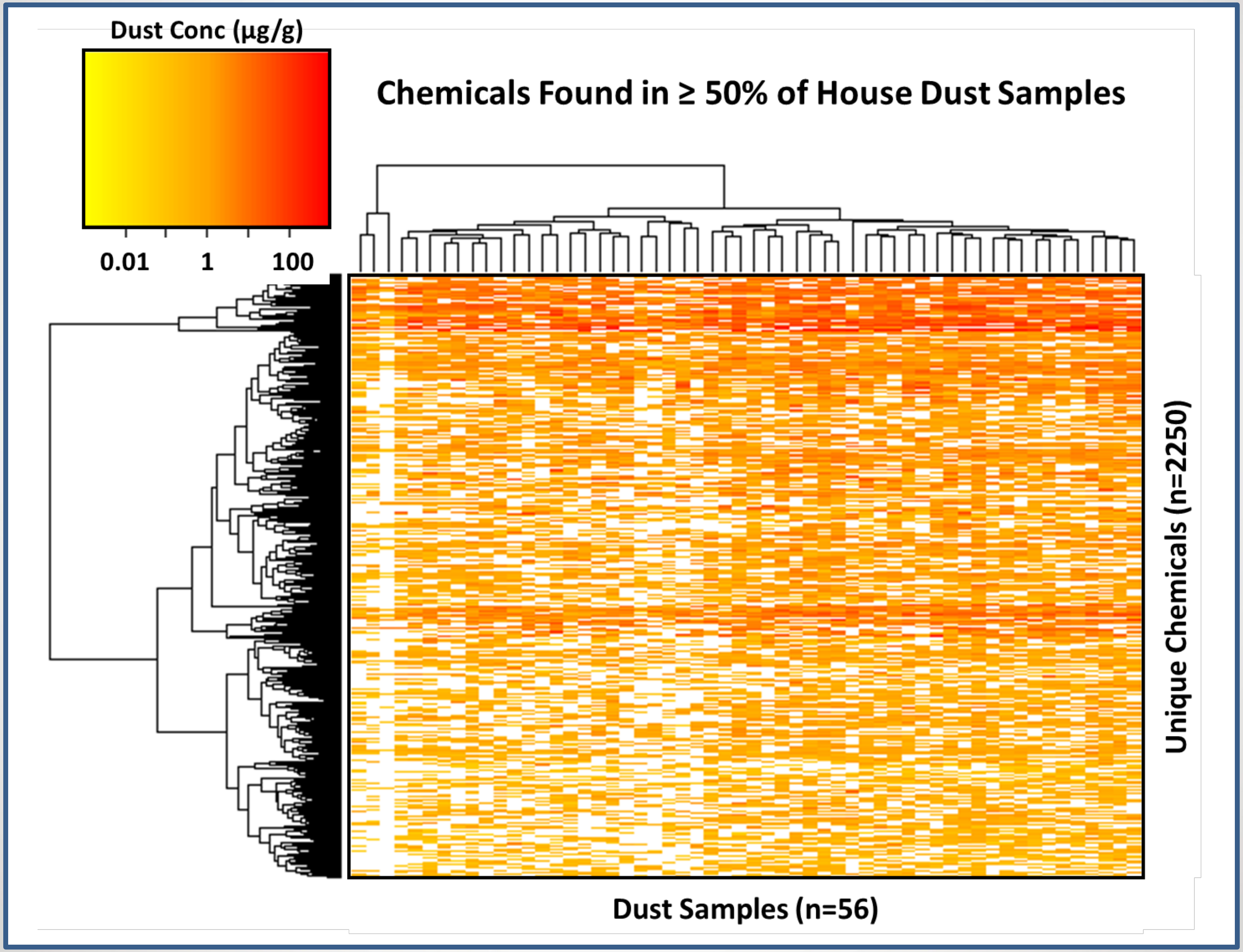




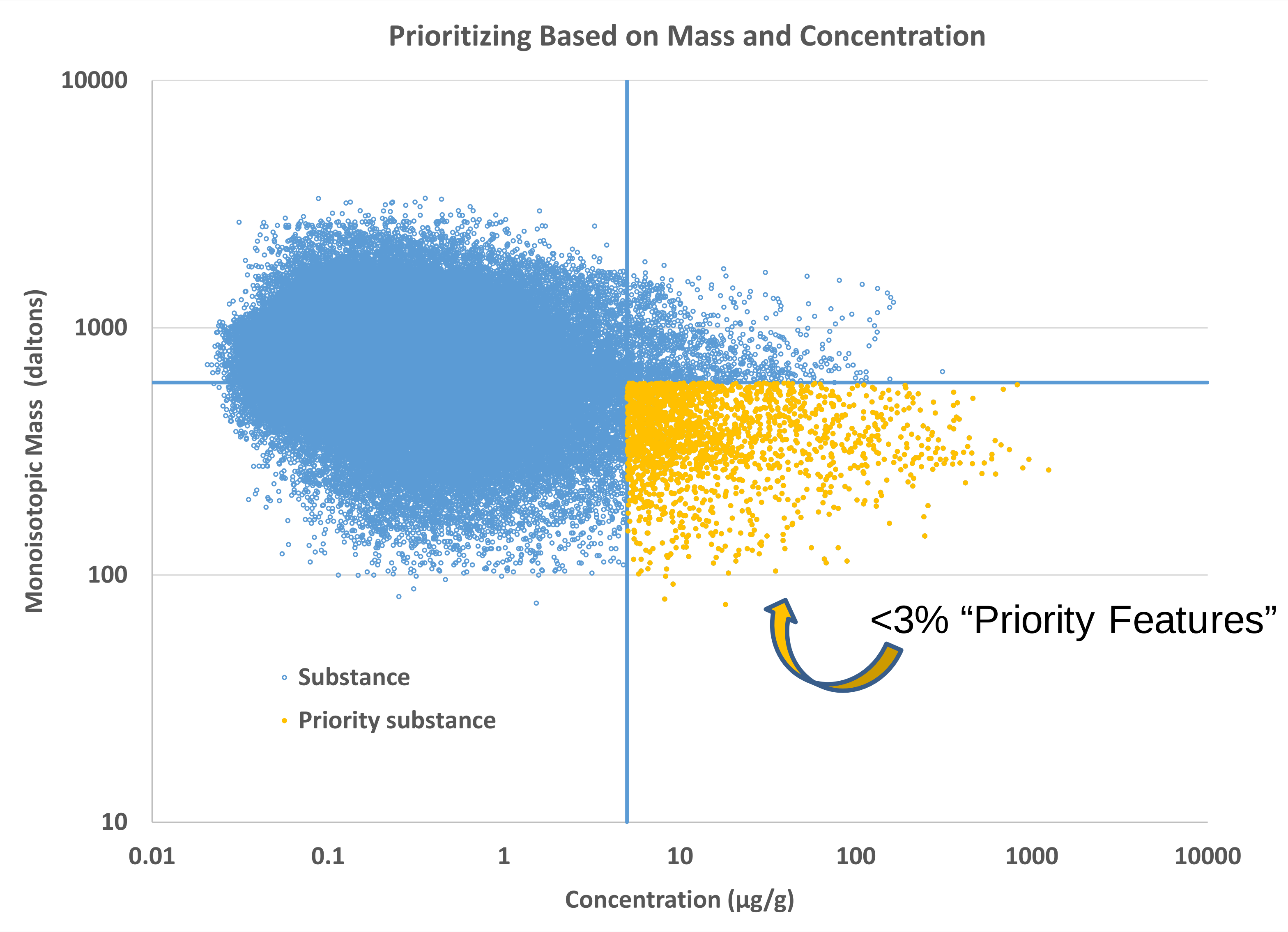
Allows conversion from peak abundance to μM units

Can convert to medium-specific units using estimated extraction efficiency

Concentration Estimates for all Features

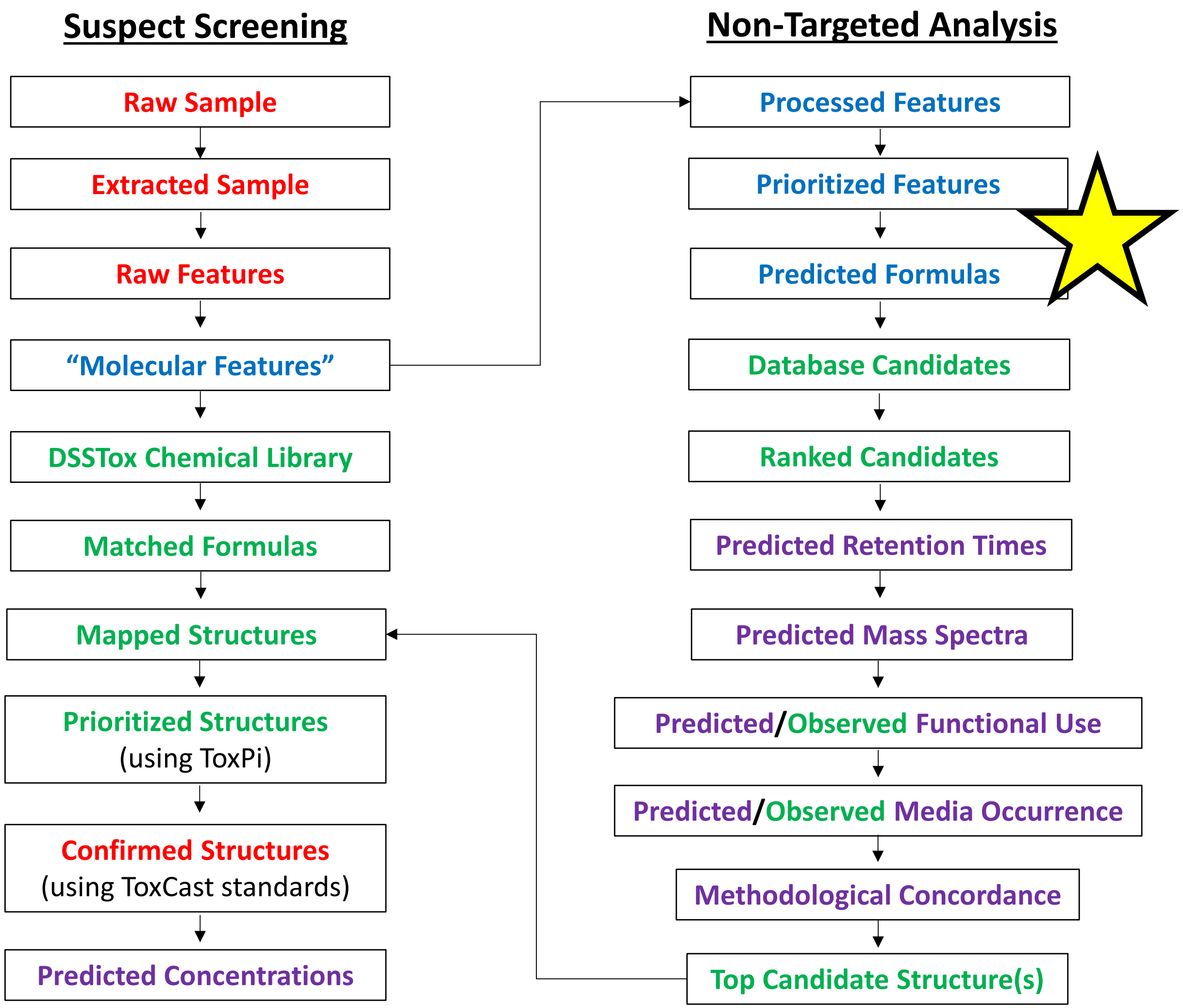


Using Mass and Concentration Filters



Material from Paul Price

Statistical Analyses for Feature Prioritization



Color Key

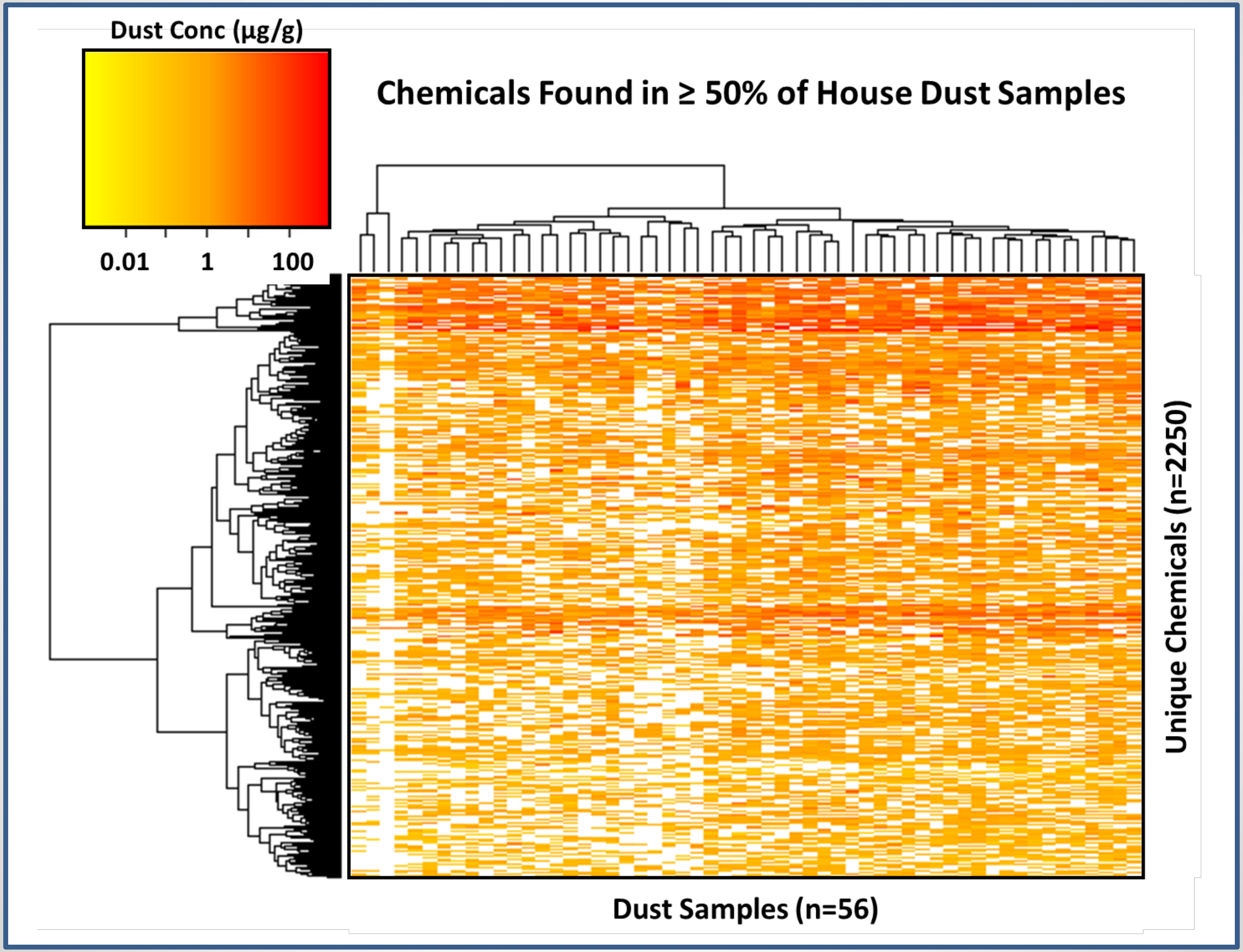
Red = Analytical Chemistry

Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

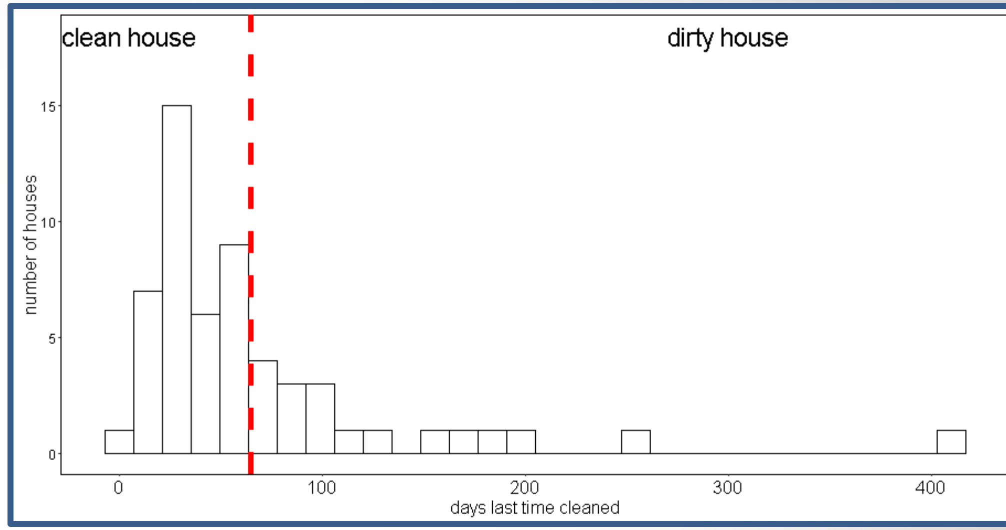
Green = Informatics & Web Services

Hierarchical Clustering



Borrowing from GWAS to Perform EWAS

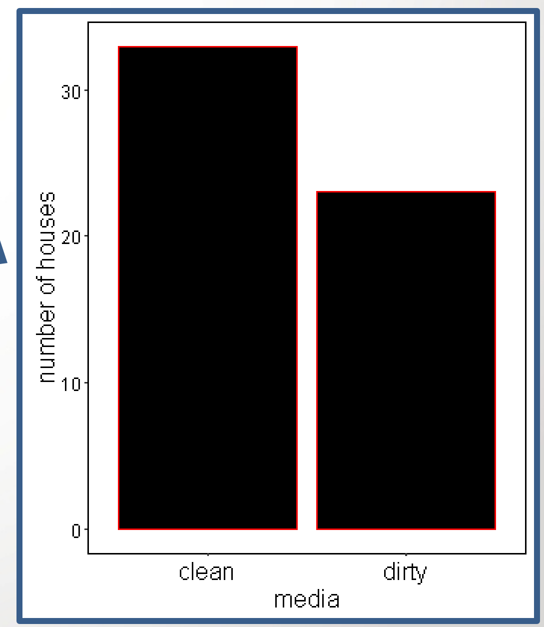
Step 1: Characterize Sources



Year Built?

Cleaning Habits?

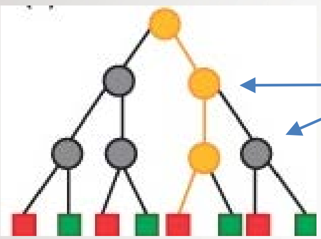
Smoking?



Material from
Derya Biryol and
Kristin Isaacs

Borrowing from GWAS to Perform EWAS

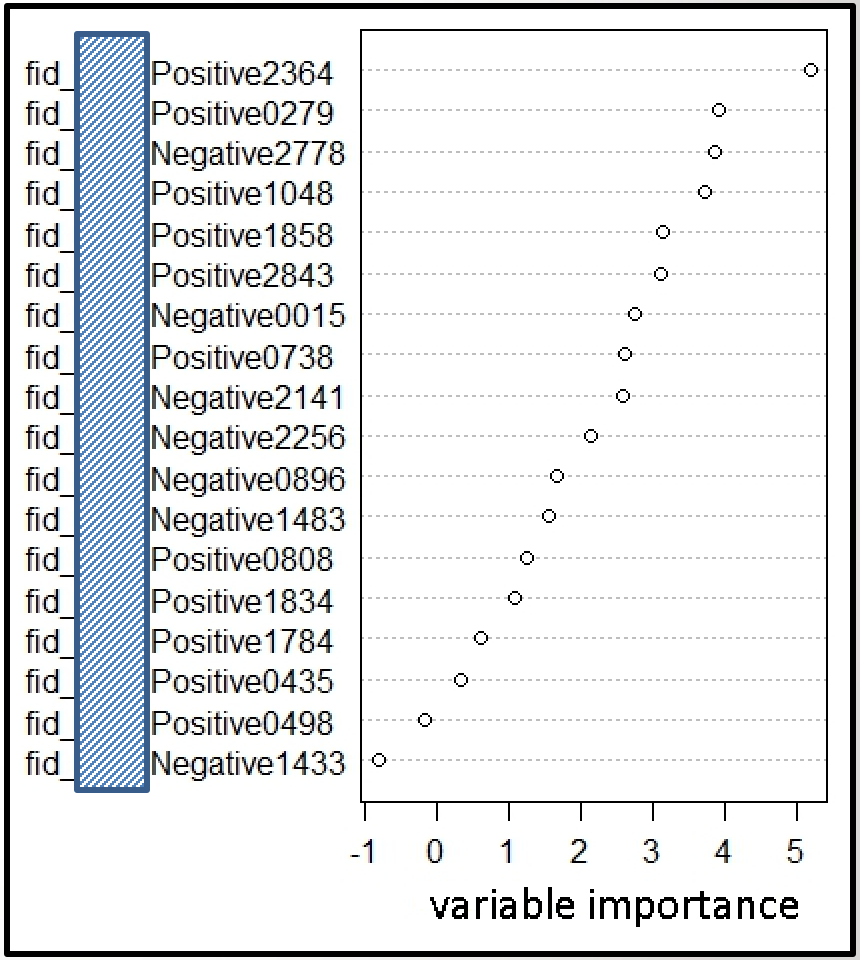
Step 2: Machine Learning Classification Modeling



Mol. Features

Exposure Classification

18 Features Associated with Cleanliness



Score	Top Predicted Formula	Monoisotopic Mass
99.52	C24 H47 N5 O	421.3756
99.43	C12 H17 N O **DEET**	191.1311
98.98	C19 H37 N8 O4	441.2947
98.1	C10 H32 N9 O3 P	357.236
97.83	C34 H63 F6 N3 O5	707.4651
97.02	C38 H84 F3 N11 O2 P2 S	877.5998
96.89	C13 H17 F N O3	254.1191
95.5	C9 H30 F N13 O P Si2	442.2002
92.82	C15 H24 F2 N O8	384.1482

Using Public Databases for Structure ID

Suspect Screening

- Raw Sample
- Extracted Sample
- Raw Features
- "Molecular Features"
- DSSTox Chemical Library
- Matched Formulas
- Mapped Structures
- Prioritized Structures (using ToxPi)
- Confirmed Structures (using ToxCast standards)
- Predicted Concentrations

Non-Targeted Analysis

- Processed Features
- Prioritized Features
- Predicted Formulas
- Database Candidates
- Ranked Candidates
- Predicted Retention Times
- Predicted Mass Spectra
- Predicted/Observed Functional Use
- Predicted/Observed Media Occurrence
- Methodological Concordance
- Top Candidate Structure(s)

Color Key

Red = Analytical Chemistry

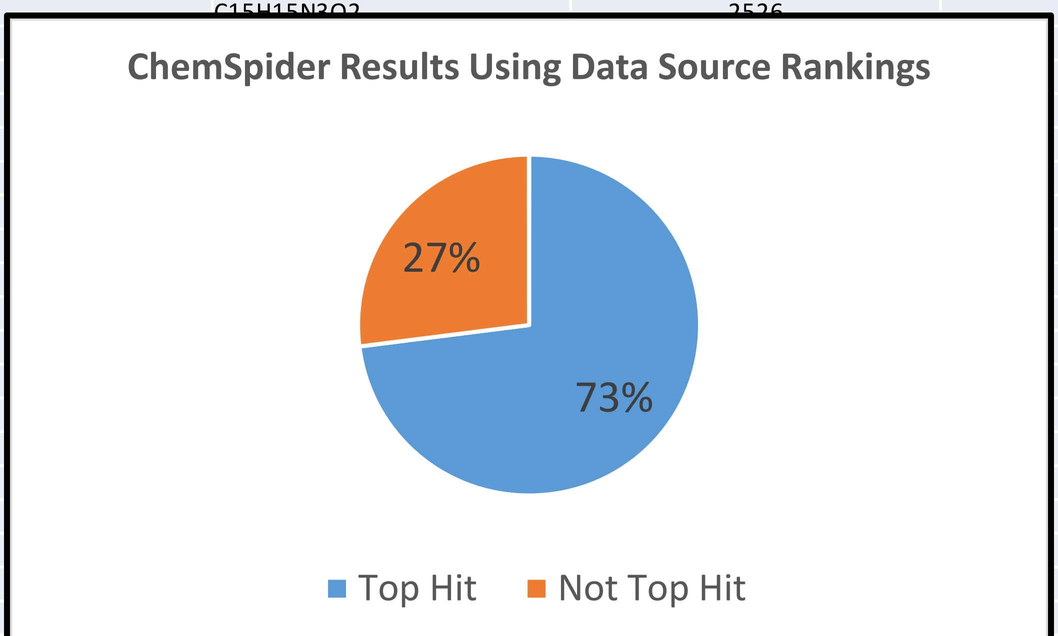
Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

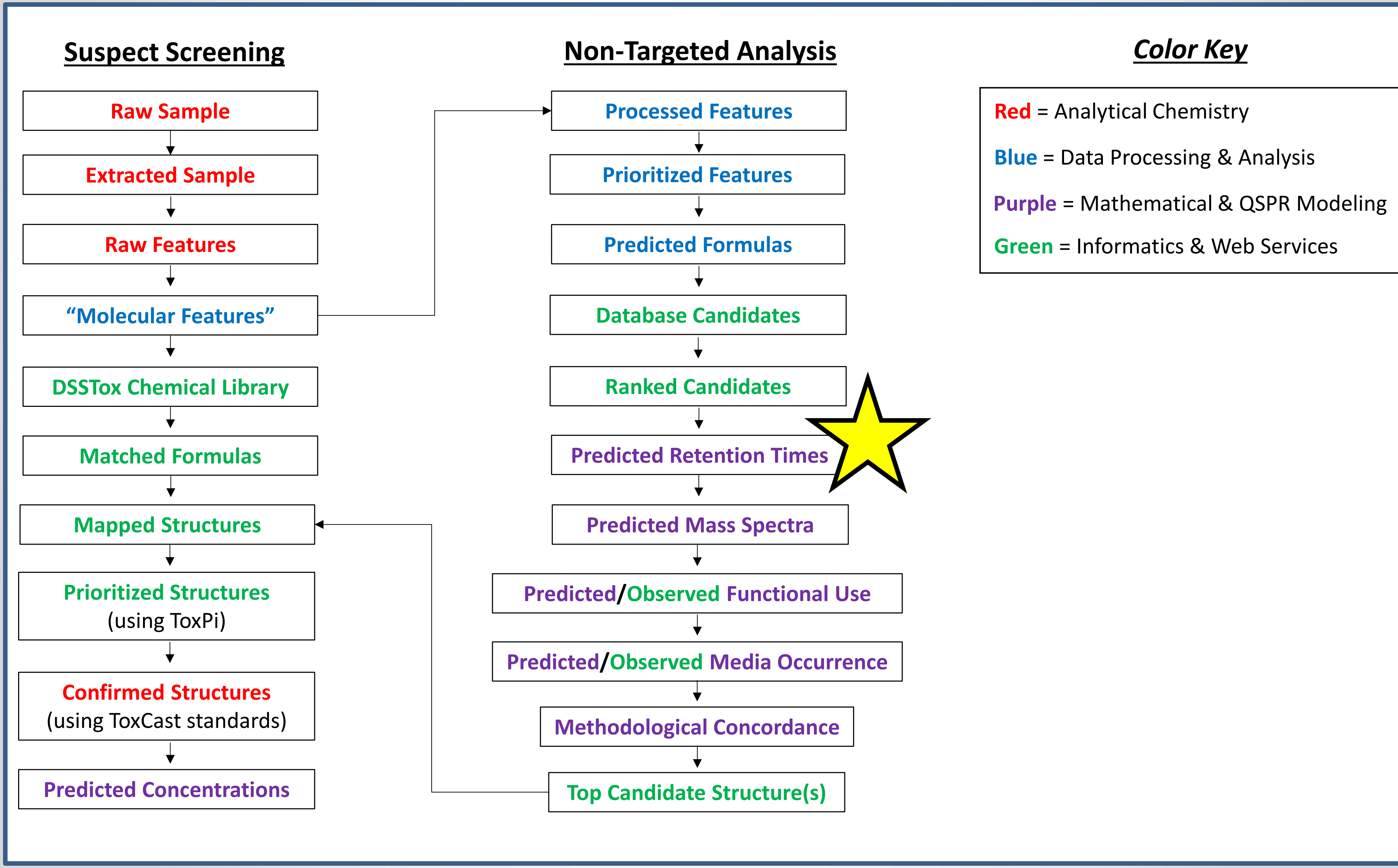
Green = Informatics & Web Services

Results for 33 Confirmed Dust Chemicals

Chemical Name	Molecular Formula	Number of Compounds with Matching Formula	Position in Results Set	Data Source Ratio
2,4,5-Trichlorobenzenesulfonic acid	C6H3Cl3O3S	12	3	0.74
2-[2-(2-Butoxyethoxy)ethoxy]ethanol	C10H22O4	59	1	1
3,6,9,12-Tetraoxahexadecan-1-ol	C12H26O5	18	3	0.83
4,4'-Sulfonyldiphenol	C12H10O4S	82	1	1
C.I. Disperse Yellow 3	C15H15N3O2	2526	3	0.38
Carbamazepine			1	1
Carbaryl			1	1
Clorophene			1	1
Corticosterone			1	1
Di(propylene glycol) dibenzoate			2	0.70
Dibutyl hexanedioate			3	0.72
Diethyl phthalate (DEP)			1	1
Diphenyl phosphate			1	1
Fluconazole			1	1
Lufenuron			1	1
Methylparaben			5	0.94
N,N-diethyl-m-toluamide (DEET)			2	0.99
N-Dodecanoyl-N-methylglycine			1	1
Nicotine			3	0.78
Octyl beta-D-glucopyranoside			1	1
Perfluorodecanoic acid (PFDA)			1	1
Perfluoroctylsulfonamide (PFOSA)			1	1
Perfluorooctanoic acid (PFOA)			1	1
Phosphoric acid, dibutyl ester	C8H19O4P	34	1	1
Piperine	C17H19NO3	3227	1	1
Primidone	C12H14N2O2	2184	1	1
Propylparaben	C10H12O3	1103	2	0.97
Rofecoxib	C17H14O4S	142	1	1
Tetradecanoic acid, 2,3-dihydroxypropyl ester	C17H34O4	47	1	1
Triclocarban	C13H9Cl3N2O	119	1	1
Triethyl citrate	C12H20O7	89	1	1
Tris(1,3-dichloro-2-propyl) phosphate (TDCPP)	C9H15Cl6O4P	8	1	1
Tris(2-ethylhexyl) phosphate (TEHP)	C24H51O4P	15	1	1



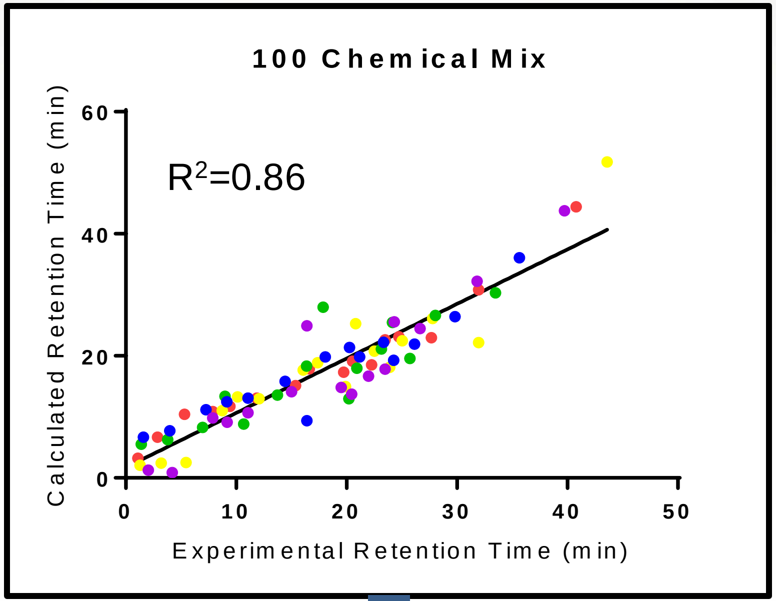
Developing/Utilizing RT Prediction Models



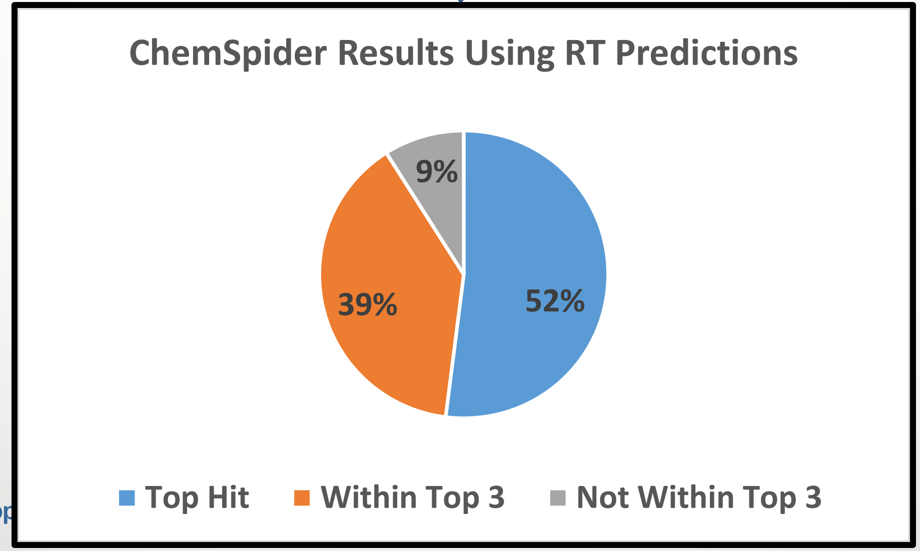
Using RT Predictions to Sort Candidates

Advanced Chemistry Development

ACD/Labs

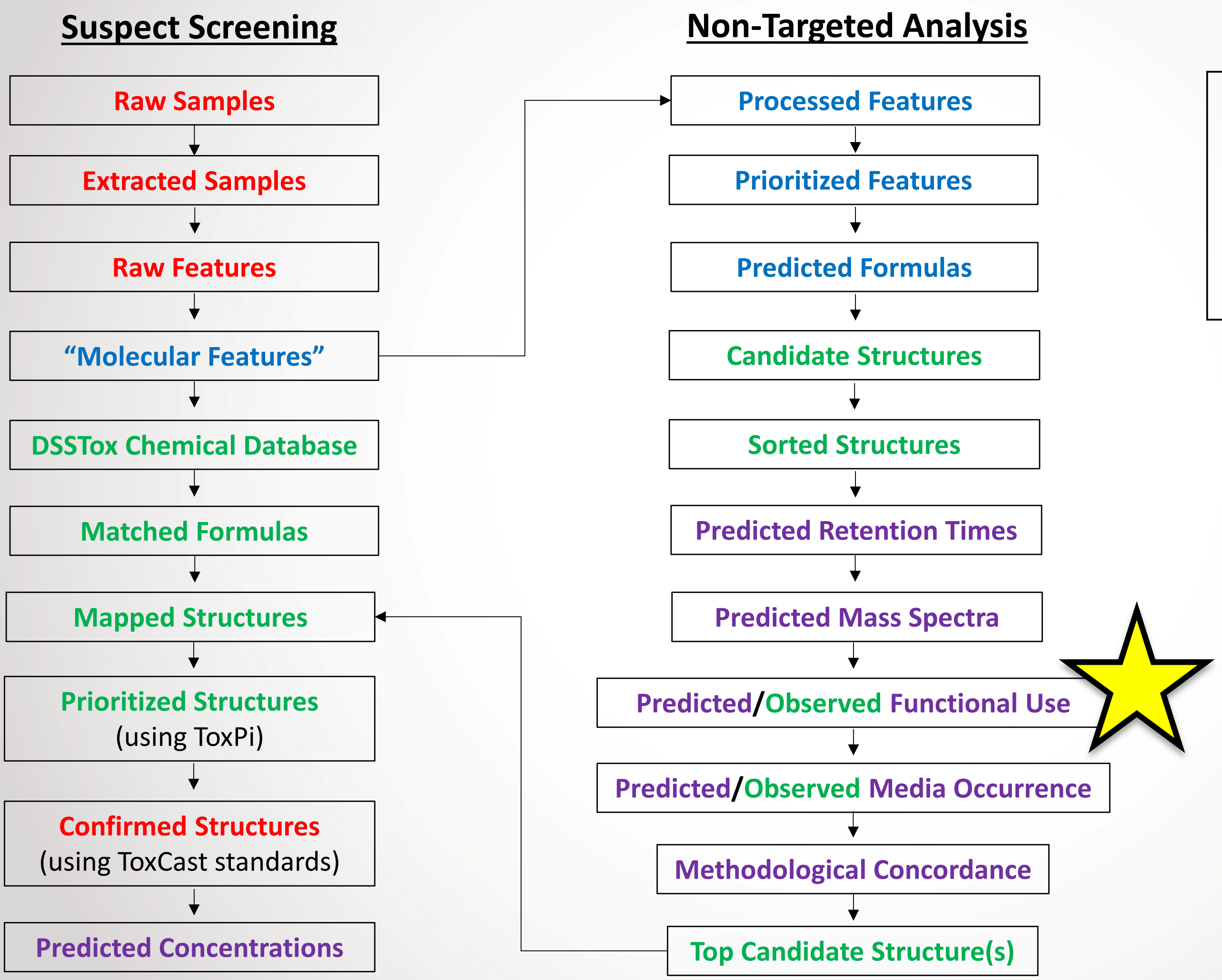


5-fold QSAR modeling approach



Material from Brandy Beverly

Utilizing Functional Use Data/Predictions



Color Key

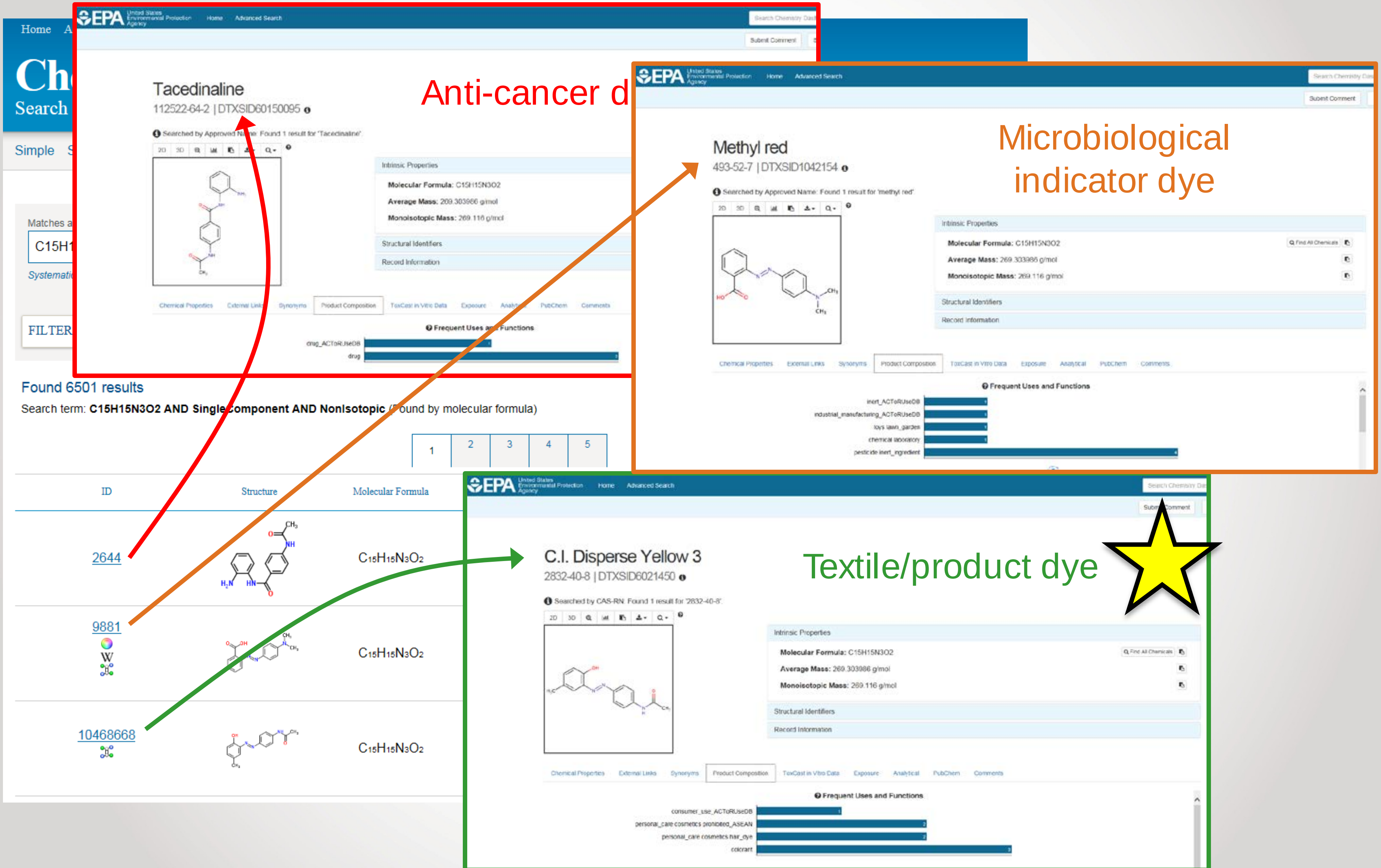
Red = Analytical Chemistry

Blue = Data Processing & Analysis

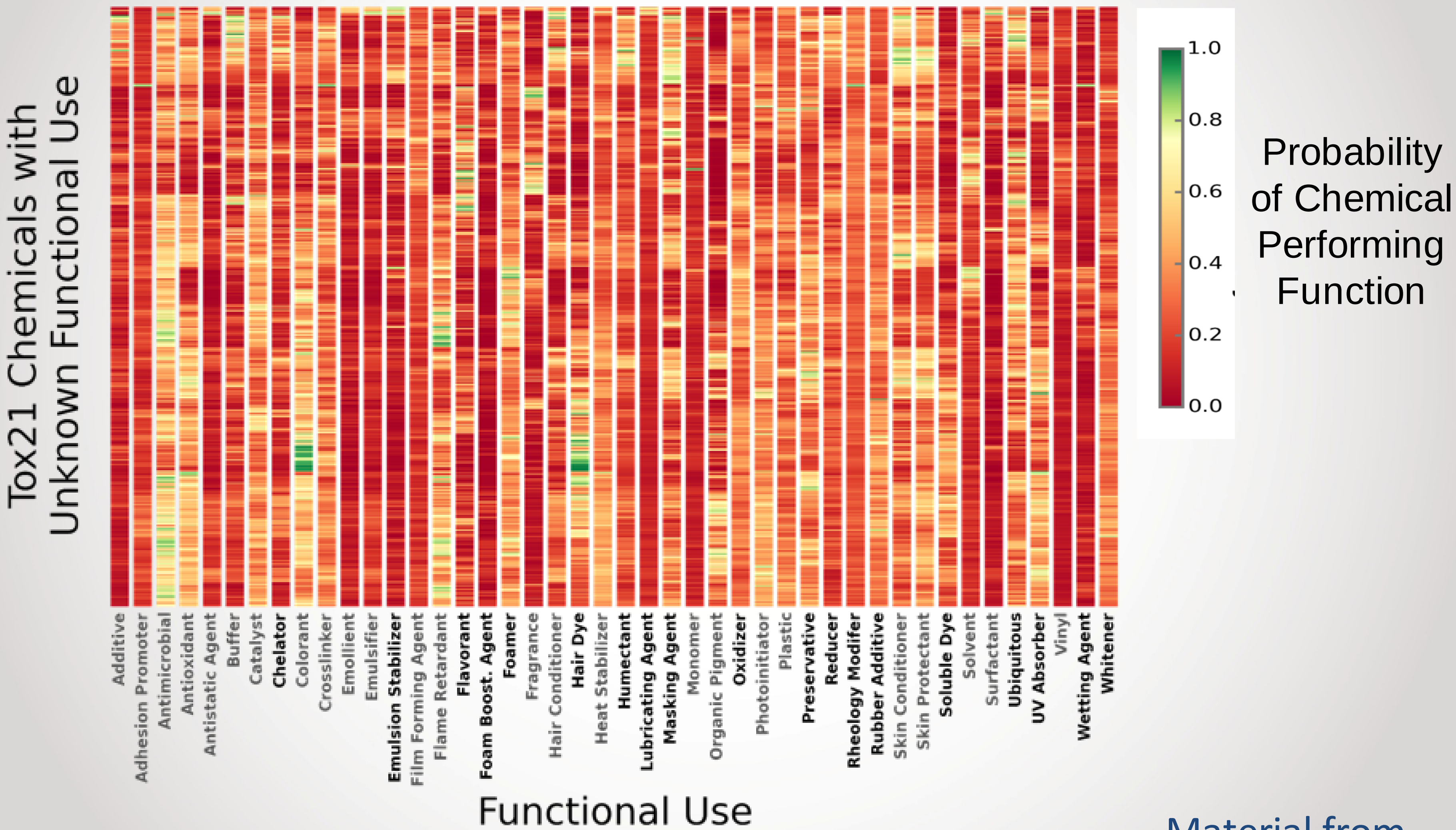
Purple = Mathematical & QSPR Modeling

Green = Informatics & Web Services

Using Functional Use to Sort Candidates

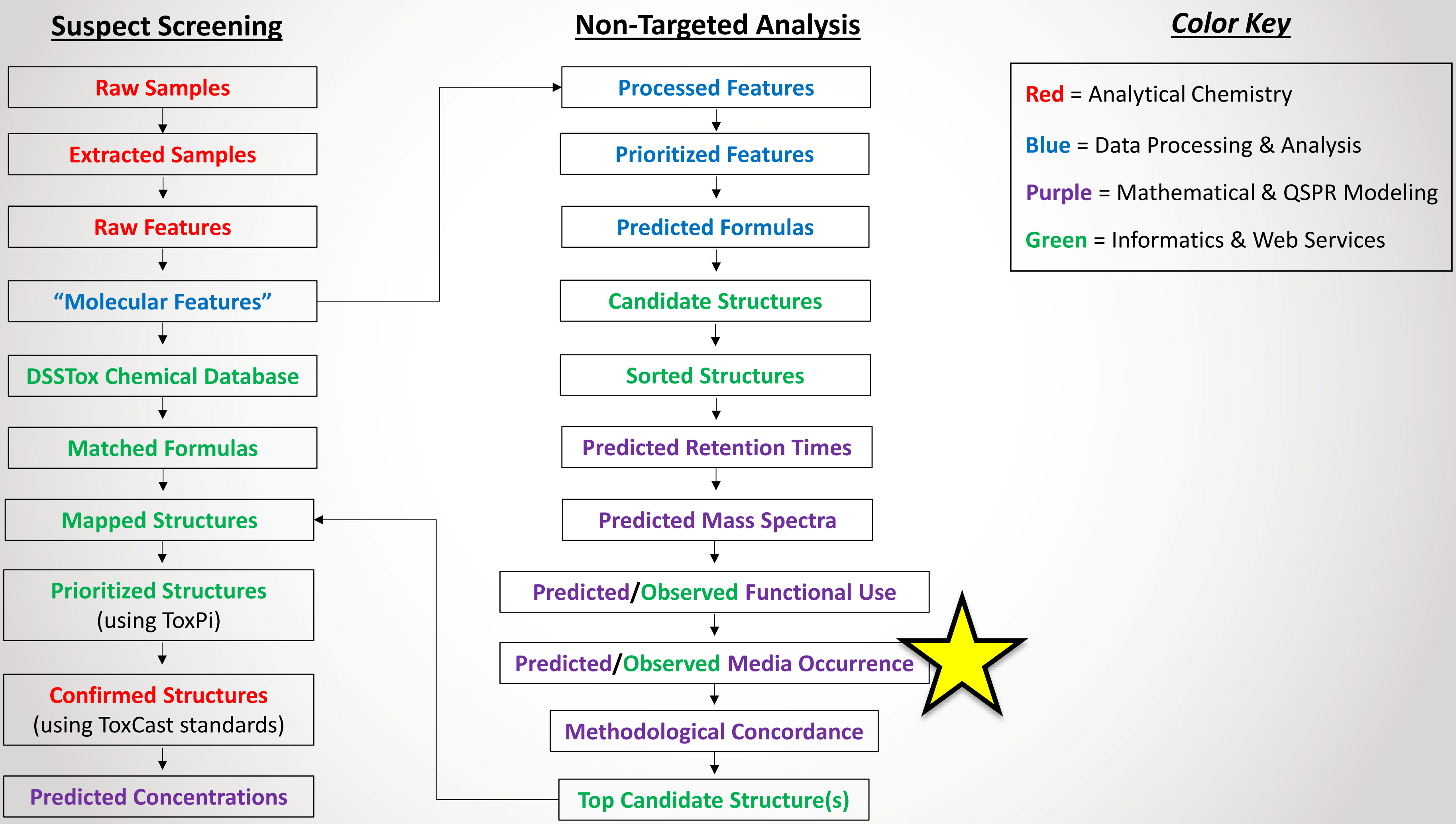


Predicting Functional Use of Chemicals



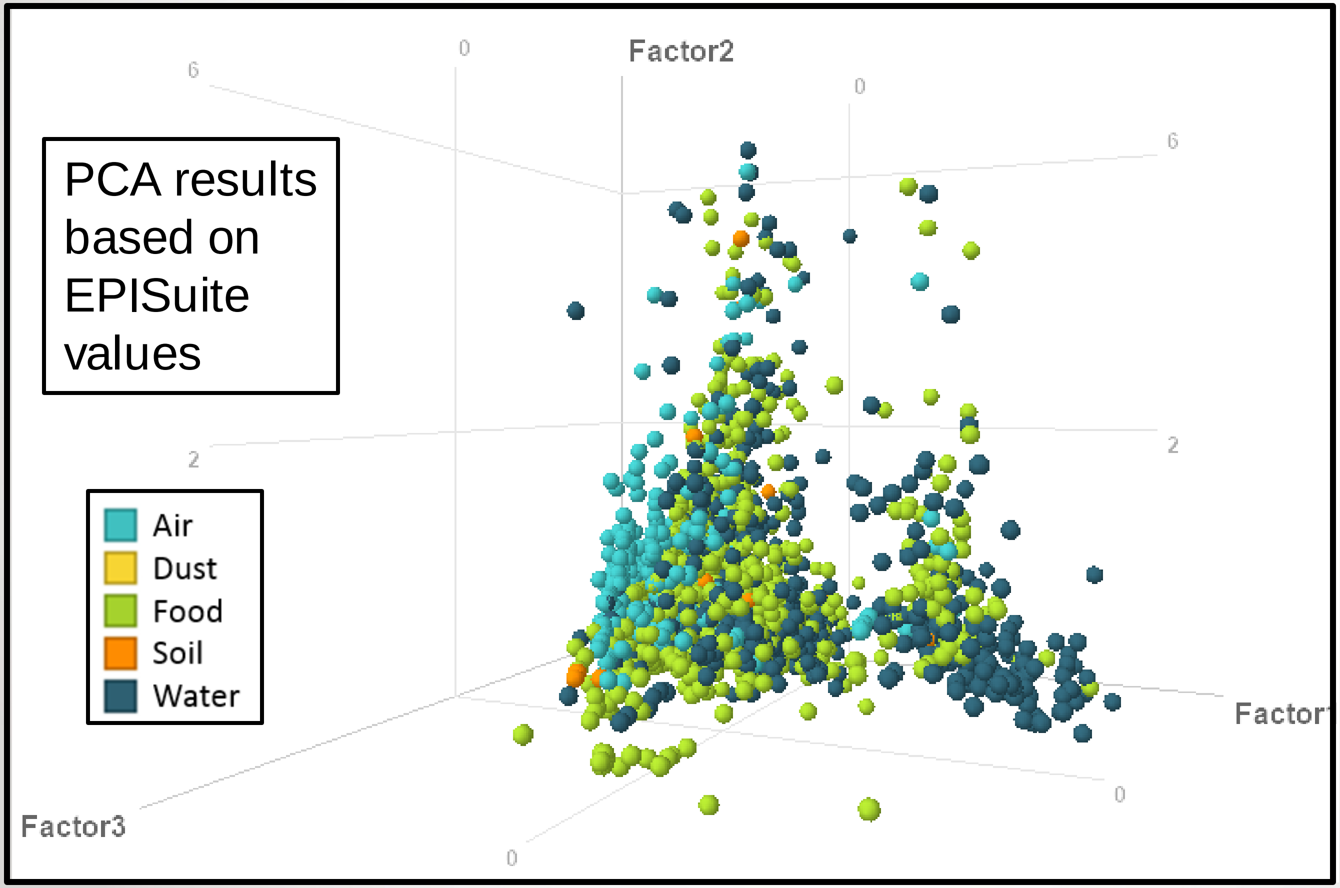
Material from Katherine Phillips

Building Media Occurrence DB & Models



Chemicals from ACToR Media

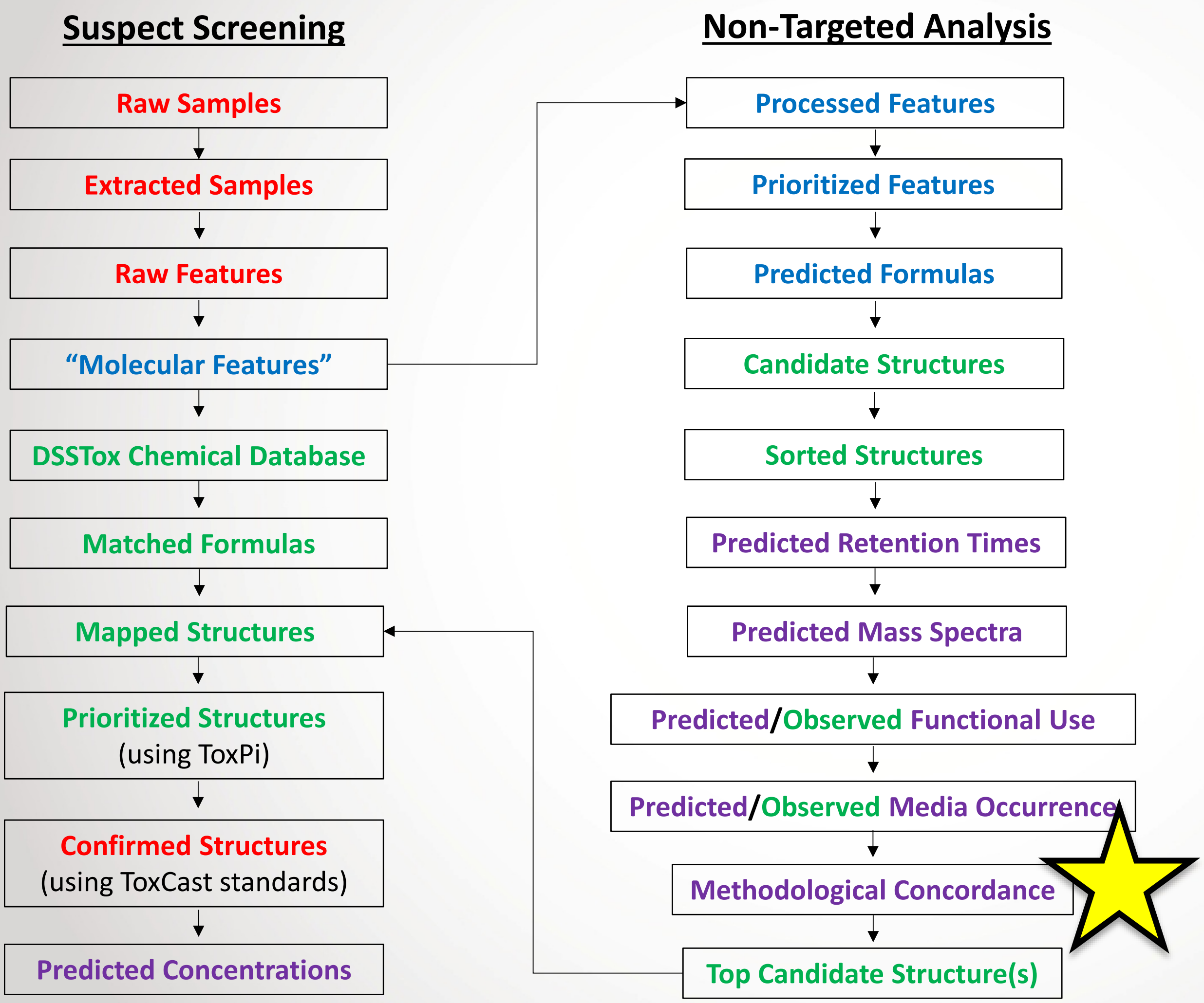
All Chemicals with Mutually Exclusive Environmental Media Categories (n=3702)



Build machine learning models based on predicted use and physicochemical descriptors

Material from Julia Rager

Finding Methodological Sweet Spots



Color Key

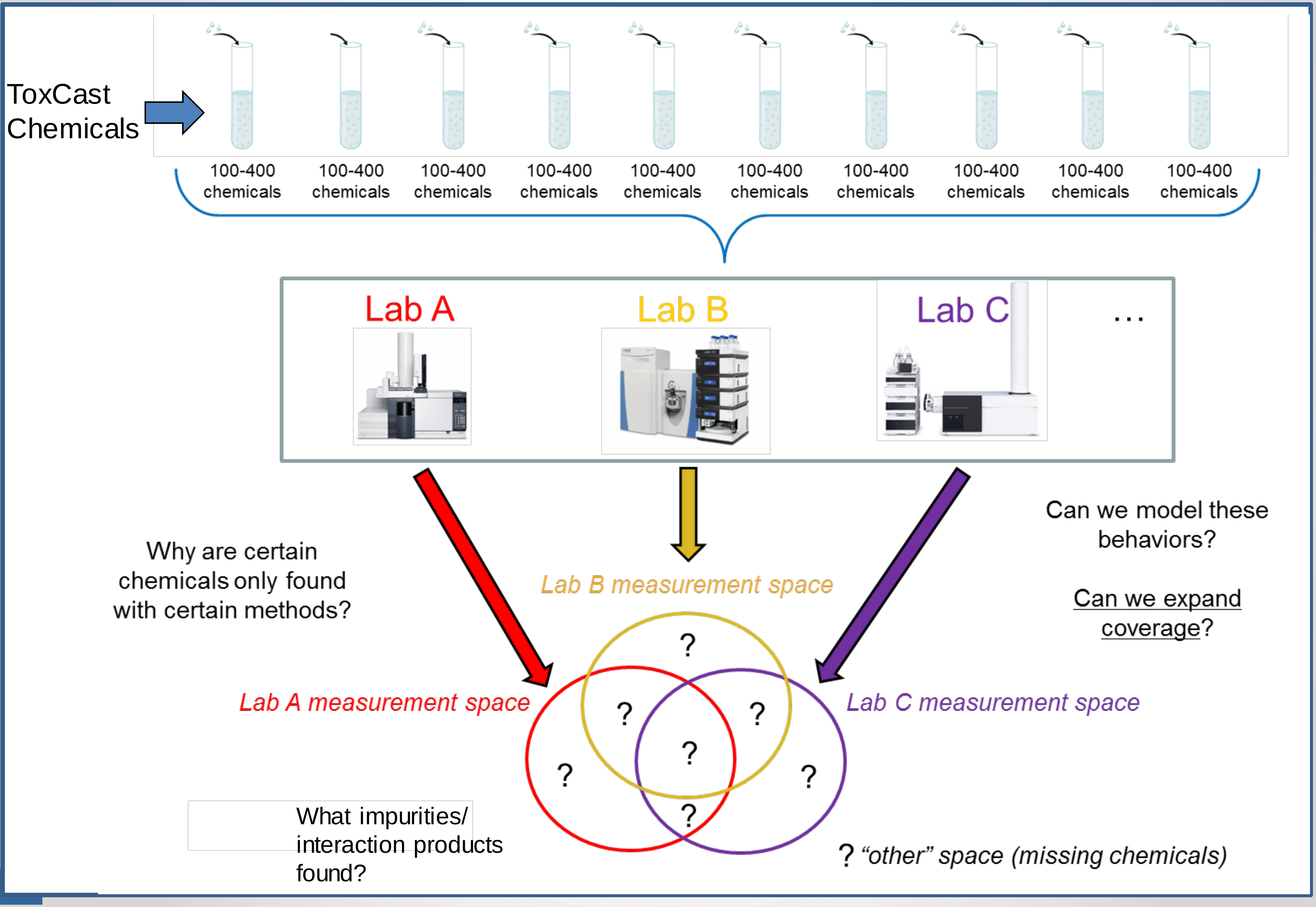
Red = Analytical Chemistry

Blue = Data Processing & Analysis

Purple = Mathematical & QSPR Modeling

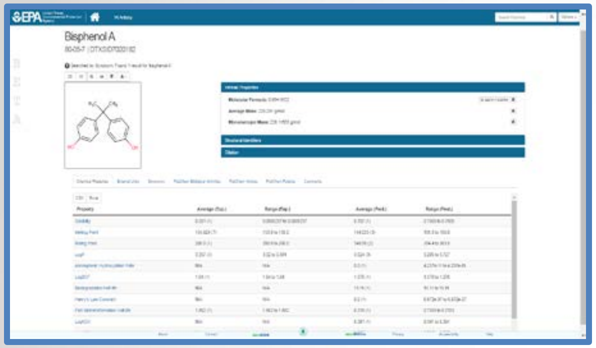
Green = Informatics & Web Services

ORD-led NTA Research Trial

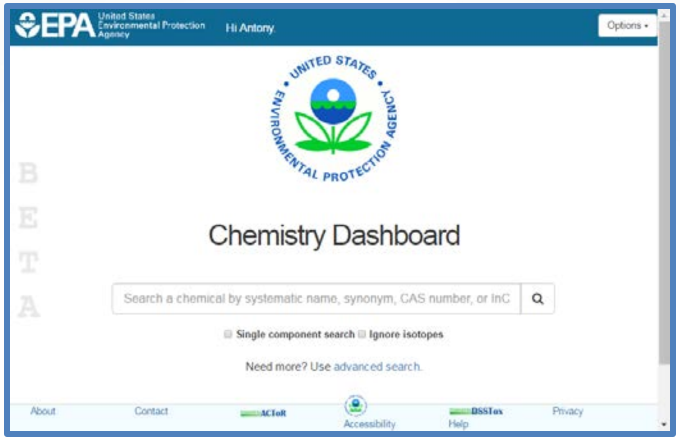


Integrating NTA Workflow Components within EPA's iCSS Chemistry Dashboard

https://comptox.epa.gov/dashboard



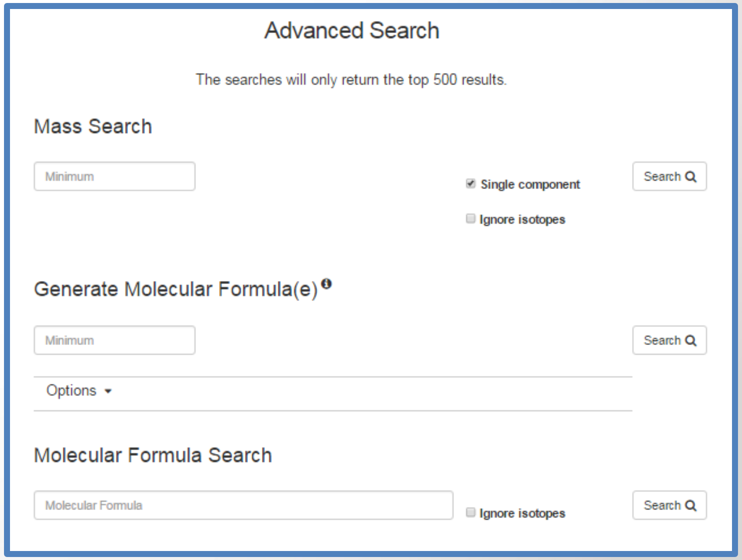
>8 million experimental and predicted physchem properties



Web access >720,000 chemicals

williams.antony@epa.gov

Integration Hub to Public Data



Advanced Searches

What About Unknown Unknowns?

Metabolites

Degradates

Even with proposed workflow, we can't find chemicals that aren't in a database

~95% of sample space often uncharacterized

Tools coming online to predict and screen for exposure dark matter

Transformation Products

Take-home Points

- ORD is developing SSA and NTA tools to support HT risk assessment
  - Applying to house dust, water/filters, silicone wristbands, serum
- Within 1 year, able to confirm up to 1300 ToxCast chemicals in media
  - ~30 laboratories (with 5 vendors) participating in NTA research trial
- New procedures being utilized to expand beyond SSA and into NTA
  - Utilizing new RT, functional-use, and media occurrence models
- New procedures required to explore "dark matter" of the exposome
  - Predictive models and workflows coming soon…

Acknowledgements

Chemical Safety for Sustainability (CSS)
Rapid Exposure and Dosimetry (RED) Project

EPA NERL
Derya Biryol*
Kathie Dionisio
Kristin Isaacs
Rebecca McMahen*
Seth Newton
Katherine Phillips
Paul Price
Jon Sobus
Mark Strynar
Elin Ulrich

EPA NCCT
Chris Grulke
Richard Judson
Kamel Mansouri*
Andrew McEachran*
Ann Richard
John Wambaugh
Antony Williams

Julia Rager
(ToxStrategies Inc.)

Brandy Beverly
(EPA NCEA)

* = ORISE Participant

Web Art Links

- Forrest vs. Trees: http://tobininvestmentplanning.com/wp-content/uploads/2015/09/do-you-see-forest-or-trees.jpg
- Black Pepper: http://blog.econugenics.com/wp-content/uploads/2014/07/blackpepper_blog_headerimage_featuredarticle-670x443.jpg
- Mad Scientist: https://upload.wikimedia.org/wikipedia/commons/thumb/9/9b/Mad_scientist_transparent_background.svg/513px-Mad_scientist_transparent_background.svg.png
- Brita Filter: https://www.brita.com/wp-content/uploads/faucet-hero1.png
- Soil in Hands: https://contentzone-bonnieplants1.netdna-ssl.com/wp-content/uploads/2011/12/soil-in-hands.jpg
- Soccer Field: http://www.ceh.org/wp-content/uploads/turf-graphic2.jpg
- Dust: http://cdn.skim.gs/images/fncsxggrflcio0qibeud/get-rid-of-dust-in-your-house
- Wastewater Effluent: http://nts-industrie.com/wp-content/uploads/sites/2/2015/09/photo-traitement-de-leaux4-200x300.jpg
- Consumer Products: http://www.findpaidfocusgroup.com/sites/default/files/CONSUMER-PRODUCTS.jpg
- Cartoon House: http://www.how-to-draw-cartoons-online.com/image-files/cartoon_house.gif.pagespeed.ce.7s_pYaegFO.gif
- Cleaning Supplies: http://www.newcf.net/wp-content/uploads/2014/03/Cleaning-supplies-1al6xdr.jpg
- No Smoking: http://a.dryicons.com/images/icon_sets/travel_and_tourism_part_1/png/512x512/no_smoking.png
- 1960: http://linabobarditogether.com/wp-content/uploads/2012/08/Year1960.png
- Decision Tree: https://www.researchgate.net/profile/John_Mitchell2/publication/260436143/figure/fig3/AS:267606825369608@1440813847562/Figure-2-Five-illustrative-decision-trees-forming-a-very-small-Random-Forest-for.png
- Dark Matter: http://7-themes.com/6797818-hd-space-wallpapers.html